

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 Введение

Одна из распространенных задач машинного обучения - получение объектов и понятий из текстов. Общая задача - научить машину понимать естественную речь. Этим занимается 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 отношений между ними.

Одной из классических задач 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. Онтологический объект - смысловая составляющая, которую несут слова в языке. Соответственно, главной целью будет 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 *Geografic reservoir*. Полезно ввести связь между более и менее общими понятиями – гипонимами и гиперонимами. Например, *Geographic location* является гиперонимом к *Geografic reservoir*. По этим связям можно выстроить многоуровневую структуру онтологических объектов, например *terrier* - *dog* - *animal*

Видно, что есть свобода в выборе уровня, к которому отнести рассматриваемый объект. Итак, основная цель - провести исследование зависимости качества получения информации от способа выбора уровня. Для этого создаётся несколько датасетов с разметкой, обозначающей одни и те же онтологические объекты метками разного уровня гипонимии. Полученное множество датасетов уже позволяет искать исследуемую зависимость. Задача выделения 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 Постановка задачи

Для улучшения качества решения задачи выделения онтологических объектов строится набор датасетов на основе произвольных английских текстов. Сначала набираются словари гипонимов-гиперонимов. Для этого берётся несколько стартовых сущностей и для каждой на основе системы Wordnet строится словарь гипонимов путём спуска вниз. В тексте для каждой сущности строится восходящая по уровню общности цепочка гиперонимов. После набора словарей достаточного размера происходит разметка: нужно зафиксировать уровень гипонимии, то есть выбрать место в полученной цепи гиперонимов. Тогда полученные датасеты будут различаться между собой конкретным выбором уровня гипонимии для каждого из слов, сущность которых мы анализируем. Далее для каждого датасета проводится обучение одной из моделей для решения рассматриваемой задачи. При этом для этого можно использовать модели и алгоритмы, которые используются для задачи поиска именованных сущностей, так как эта задача похожа на решаемую здесь. Тогда, используя для выбранной модели соответствующую метрику качества, можно определить наиболее эффективные по выделению онтологических объектов уровни гипонимии. В частности, предлагается воспользоваться SOTA алгоритмам из [6]. При этом различать датасеты можно по-разному, можно различать по изменению одной метки, а можно по увеличению уровня гипонимии на всех словах. В первом случае разницы между датасетами почти не будет, а нужны разные. Во втором же возникает проблема разметки нового датасета, когда для разных слов цепочки гиперонимов будут иметь разную длину.

3 Список литературы

Методы решения задач распознавания именованных сущностей активно развиваются, начиная с девяностых годов прошлого века. Одной из первых была работа [1], где предлагался эвристический подход к решению задачи. Дальнейшее развитие методов в течение десяти лет хорошо описывается в [2]. Далее исследования продолжаются, часто уходя в узкие области для улучшения эффективности в частных случаях. Так в [3] рассматриваются подходы, основанные на применении нейросетей, где 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 это делалось в [4].

Также хорошо показали себя модели на вероятностной основе. Например, CRF, описанная в [5], где приводятся алгоритмы выдающие вероятности возможных продолжений в заданной текстовой позиции.

Рассматриваемая задача поиска онтологических объектов является более общей, чем задача распознавания именованных сущностей, но подход к обоим задачам аналогичен

Литература

- [1] *Rau L. F.* Extracting names from text Proc. of the Seventh Conference on Artificial Intelligence 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] *Sheng E., Natarajan P.*, A Byte-sized Approach to Named Entity Recognition www.arxiv.org/pdf/1809.08386v1.pdf
- [4] *Cao S.* Deformable Stacked Structure for Named Entity Recognition www.arxiv.org/pdf/1809.08730v2.pdf
- [5] *Lafferty J.* Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
<https://arxiv.org/pdf/1709.09686.pdf>
- [6] Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition