Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

Дочкина В., Кузнецов М., Резяпкин В., Русскин А., Ярмошик Д.
snikyu@gmail.com

В данной работе исследуется задача определения онтологических сущностей в тексте. Собран датасет гипонимий с использованием ресурса Wordnet. Данный датасет применяется на различных текстовых корпусах для автоматической разметки слов. На данной разметке исследуется качества работы Named Entity Recognition алгоритмов. Одним из приложений результатов данной работы может быть добавление новых признаков слов в задачах Nature Language Processing для повышения качества

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

1 Введение

Одним из крупных разделов машинного обучения является Nature Language Processing – обработка естественного языка. Диалоговые системы, машинный перевод, анализ тональности, моделирование языка и т.д. - лишь небольшое подмножество задач NLP. Из этого множества задач выделяется подмножество, называемое named entity recognition, или распознавание именованных сущностей. Основной задачей является распознавание и разметка слов и словосочетаний, являющихся именами собственными в тексте. Например, текст “Радько родился в 1957 году в Ленинграде” можно разметить как “$[\text{Радько}]_{\text{личность}}$ родился в 1957 году в $[\text{Ленинграде}]_{\text{город}}$” .

Мы рассмотрим некоторое обобщение данной задачи. А именно, помимо именованных сущностей будем распознавать онтологические объекты. Для примера возьмём слово "цветок". В задаче распознавания именновах сущностей данному слову не будет назначена метка. В случае же работы с онтологическими объектами это слово будет размечено. Более того – не одну. Возьмём еще три слова: “роза”, “растение” и “живой организм”. Ясно, что роза является цветком, цветок – растением, а растение – живым организмом. Связи такого рода, частное-общее, называются гипонимией. Если слово A является частным слова B, то говорят, что A является гипонимом B, и, в обратную сторону, B является гиперонимом A. Например, “цветок” – гипоним “растения” и гипероним “розы”.

Таким образом, нашей задачей является разметка текста на различных уровнях гипонимии. Так мы получим несколько датасетов с разными разметками онтологических объектов.

Задача является, актуальной потому что ... (результаты можно использовать для обучения других моделей)

Датасет будет получен с помощью словаря гипонимий электронного тезауруса Wordnet. На полученных данных будем исследовано, как зависит качество распознавания онтологических объектов от глубины гипонимии. Кроме того, есть желание настроить модель так, чтобы она давала метки новым для неё словам. Это возможно осуществить с помощью учёта контекста и поиска похожих слов. Например, можно давать аналогичные метки словам с почти равными векторными представлениями.

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

В первую очередь, необходимо определить и извлечь множество онтологических объектов с заданной глубиной гипонимии с электронного тезауруса Wordnet. При составлении данного множества преследуется задача получения в некотором смысле близких по обобщенности онтологических объектов на одинаковых уровнях. Полагая единственность гиперонима для каждого объекта, структуру системы гипонимов онтологического объекта удобно принимать деревом. Таким образом, множество меток задачи хранится в виде набора деревьев с равной глубиной h.

Следующим этапом является разметка датасетов. Каждый датасет необходимо разметить на различных уровнях глубины гипонимии. Так как, для объектов верхнего уровня гипонимии гиперонимы отстутствуют, количество разметок равняется $h - 1$.

В заключение, на размеченных данных необходимо провести несколько экспериментов с использованием SOTA NER алгоритмов. Полученные результаты будут проанализированы на зависимость качества распознавания онтологических объектов в зависимости от глубины гипонимии.

3 Обзор литературы

За последние несколько лет задача распознавания именованных сущностей решалась с помощью нескольких подходов и фреймворков [1, 2]. Существуют подходы, использующие сгенерированные "руками"признаки, такие как части речи, предыдущие и последующие слова, а также признаки, порожденные регулярными выражениями. Основным недостатком является производительность таких моделей в зашумленных данных, например, твитах.

Последние модели использует нейросетевые подходы для решения задачи [3]. Аналогично с точки зрения архитектуры, в [4] также предложена модель, которая учится на векторных представлениях слов, кодирующих семантические и синтаксические признаки слов для различных естественных языков языке. В [5] используется нейронная сеть, которая автоматически обнаруживает признаки слов и символов с помощью комбинации BiLSTM и CNN.

Современные подходы к решению задачи Named Entity Recognition, разработанные в течение последних нескольких лет, используют комбинации нейронных сетей и Conditional Random Fields. Такие подходы дают высокую точность распознавания именнованных сущностей. К примеру, в [6] используется модель, пропускающая векторные представления знаков и слов через BLSTM слой, за которым следует CRF слой. Модель BiLSTM+CRF дала высокие результаты на русскоязычных датасетах [7]. Эти модели также успешно применяются на биомедицинских датасетах [8, 9].

Литература

[1] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. 2007.

[2] Angus Roberts, Robert J. Gaizauskas, Mark Hepple, and Yikun Guo. Combining terminology resources and statistical methods for entity recognition: an evaluation. 2008.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. 2016.

[4] Rami Al-Rfou', Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Polyglot-ner: Massive multilingual named entity recognition. 2015.

[5] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *TACL*, 4:357–370, 2016.

[6] Xuezhe Ma and Eduard H. Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354, 2016.

[7] L. T. Anh, M. Y. Arkhipov, and M. S. Burtsev. Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition. *CoRR*, abs/1709.09686, 2017.

[8] Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. Long short-term memory rnn for biomedical named entity recognition. 2017.

[9] Mourad Gridach. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, 70:85–91, 2017.