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 Описание основных методов

Для задачи NER используется модель нейронной сети с гибридной архитектурой: Bi-LSTM-CRF.

3.1 Рекуррентные нейронные сети

Для учета контекста в тексте используются рекуррентные неиронные сети (RNN). В отличие от многослойных перцептронов, рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины. В реальности обычная RNN хранит информацию только о коротком контексте (затухание градиентов). Такого недостатка лишена LSTM – неиросетевои рекуррентныи блок, состоящии из элементов: Основнои слои (как и в обычнои RNN), три сигмоидальных слоя-фильтра, Ячейка памяти. Формулы для этих компонент:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f), \tag{2}$$

$$c_n = g(W_{cx}x_t + W_{ch}h_{t-1} + b_c), \tag{3}$$

$$c_t = f_{tt-1} + i_t \circ c_n, \tag{4}$$

$$h_t = o_t \circ g(c_t), \tag{5}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{6}$$

Переменные:
x_t - входной вектор
h_t — выходной вектор,
c_t — вектор состояний,
W, Ub — матрицы параметров и вектор,
$f_t, i_t o_t$ — векторы вентилей,
f_t — вектор вентиля забывания, вес запоминания старой информации,
i_t — вектор входного вентиля, вес получения новой информации,
o_t — вектор выходного вентиля, кандидат на выход. Функции активации:

σ_g : на основе сигмоиды.
σ_c : на основе гиперболического тангенса.
σ_h: на основе гиперболического тангенса, но в работе о глазках (смотровых отверстиях) для LSTM предполагается, что $\sigma_h(x) = x$

Сигмоида :

$$f(x) = \frac{1}{1 + e^x}, \tag{7}$$

- $f(x) \in [0, 1]$ - позволяет моделировать вероятности
- Дифференцируема и монотонна
- Обощается функцией softmax

Гиперболический тангенс

$$f(x) = tanh(x) \tag{8}$$

- Все свойства сигмоиды
- Значение $f(x)$ всегда неотрицательно
- Обычно используется для бинарной классификации

3.2 Bi-LSTM

Двунаправленные рекурентные нейронные сети (Bi-LSTM) были разработаны для кодирования каждого элемента в последовательности с учетом левого и правого контекстов, что делает его одним из лучших вариантов для задачи NER. Обычная LSTM учитывает только прошлыи контекст, двунаправленная учитывает и будущии. Расчет двунаправленной модели состоит из двух этапов: первый слой вычисляет представление левого контекста и второй слой вычисляет представление правого контекста. Выходы этих шагов затем объединяются для получения полного представления элемента входной последовательности. Было показано, что двунаправленные кодеры LSTM полезны во многих задачах NLP, таких как машинный перевод, ответ на вопросы, и особенно для решения проблемы NER.

3.3 CRF модель для задачи NER

Conditional Random Field (CRF) - вероятностная модель для структурного прогнозирования, которая успешно применяется в различных областях, в том числе для обработки естественного языка.
Модель CRF обучается предсказывать вектор $\boldsymbol{y} = y0, y1, ...y_n$ тегов с учетом предложения

$\boldsymbol{x} = x0, x1, .., x_N$. Для этого вычисляется условная вероятность:

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{e^{Score(\boldsymbol{x},\boldsymbol{y})}}{\sum_{y'} e^{Score(\boldsymbol{x},\boldsymbol{y})}} \tag{9}$$

где Score рассчитывается по формуле :

$$Score(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_i y_i, \tag{10}$$

где$A_{y_i,y_{i+1}}$ обозначает вероятность перехода от тега i к тегу j, $P_{i,j}-$ вероятность перехода, которая представляет оценку j-го тега i-го слова.

3.4 комбинация Bi-LSTM и модели CRF

В комбинированной модели символы каждого слова в предложении подаются в сеть Bi-LSTM для того, чтобы охватить особенности слов на уровне символов. Затем эти векторные представления уровня символов объединяются с векторами встраивания слов и передаются в другую сеть Bi-LSTM. Эта сеть вычисляет последовательность оценок, которые представляют вероятности тегов для каждого слова в предложении. Чтобы повысить точность прогнозирования, уровень CRF обучается применять ограничения, зависящие от порядка тегов. Например, в теге схемы BIO (B - Begin, I - Inside, O - Other) I никогда не появляется в начале предложения, или O I B O - недопустимая последовательность тегов.

4 Эксперимент

В ходе эксперимента была использована система wordnet, позволяющая получить информацию о деревьях гипонимов слов. Для эксперимента был выбран гипероним entity и входящие в него три уровня гипонимов.
Например, первый уровень состоит из : physical entity и abstract entity. В качестве набора данных был использован датасет Web of sciense, состоящий из набора текстов с различными контекстами. Для разметки данных выбран формат BIO и для каждого уровня гипонимии создан отдельный датасет, разделенный на обучающую, тестовую и валидационную выборки. Для обучения использовалась описанная выше модель. Оценка качества проводится F_1 метрикой.

F-мера:
F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю. Выражение для F_1 меры:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall}, \tag{11}$$

Recall - полнота,
Precision - точность

Результаты:

Таблица 1. F_1 мера для разных уровней гипонимии

Первый уровень:	**75.09**
Второй уровень:	**67.37**
Третий уровень:	**56.92**

5 Заключение

Результаты проведенного эксперимента показывают, что с увеличением уровня гипонимии качество распознавания онтологических объектов падает.

6 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.