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 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 отношениz 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. Онтологический объект - смысловая состовляющая, 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 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. 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 соварей достаточного размера происходит разметка: нужно зафиксировать уровень гипонимиии , 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 гипонимиии.
В частности, предлагается воспользоваться SOTA алгоритмом из [6]
При этом различать датасеты можно по-разному, можно различать по изменению одной метки, а можно по увеличению уровню 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 3 Описание основных методов

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

4 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

5 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.

6 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.

7 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.

8 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:

Таблица1 F_1 мера для разных уровней гипонимии

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

9 5 заключение

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

10 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 случаяях. 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 этоделалось 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 сушностей, но подход к обеим задачам аналогичен

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] *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