Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

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

Данная работа посвящена распознаванию именованных сущностей с использованием алгоритмов машинного обучения. Исследуется зависимость качества распознавания онтологических объектов от глубины гипонимии, основанная на нейронной модели для распознавания именованных сущностей. Работа была проведена на наборах данных с использованием различных видов разметки и глубины вложенности объектов. В результате проведенного исследования была определена зависимость качества распознавания сущностей от глубины гипонимии

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

1 Введение

Распознавание именованных сущностеи (Named-entity recognition, NER) является однои из самых важных задач в области обработки естественного языка (Natural language processing, NLP). Задача NER состоит в выделении и классификации именованных сущностеи в тексте. Именованная сущность - это слово или словосочетание, обозначающее предмет или явление определеннои категории. Именованными сущностями являются личности (PERSON), названия организаций (ORGANIZATION), локации (LOCATION) и другие. Пример размеченного текста: "[ORGANIZATION MIPT], is a Russian university, located in [LOCATION Dolgoprudny]". Впервые задача извлечения именованных сущностеи была поставлена на конференции Message Understanding Conference (MUC) в 1996 году. Позднее она упоминалась на конференциях Conference on Computational Natural Language Learning (CoNLL) CoNLL-2002 и CoNLL-2003.

В даннои статье рассматривается случаи извлечения онтологических объектов при разных заданных уровнях гипонимии. Гипонимия - вид системных отношении в лексике: связь слов по линии «общее и частное», отражение родо-видовых отношении между явлениями деиствительности. Гипонимы - слова, называющие предметы (своиства, признаки) как элементы класса (множества) и состоящие в отношениях гипонимии со словом — названием этого класса (гиперонимом). Например, слова: кольцо, браслет, ожерелье являются гипонимами по отношению к слову украшение, и наоборот, с точки зрения обратного отношения, украшение выступает как гипероним по отношению к словам кольцо, браслет, ожерелье.

Целью исследования является выделение объекта и последующего определения его к конкретному типу, то есть сопоставление гипонима его гипериому. При этом глубина гипонимии может быть многоуровневои. Например "Mont Blanc"относится к классу Mountain, которыи в свою очередь относится к классу Geographic location.

Полученная информация об онтологических объектах оказывается полезнои в различных задачах, связанных с обработкои естественного языка, включая ответы на вопросы,заданные пользователем и извлечении зависимостеи между объектами. Рассмотренная в данном исследовании проблема может быть использована для производства дополни-

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

тельных признаков при решении различных NLP задач, а также определения - являются ли объекты парои гипоним-гипероним.
Один из стандартных подходов к NER заключается в рассмотрении проблемы как задачи маркировки последовательности, когда каждому слову присваивается тег, указывающии является ли слово частью именованнои сущности или появляется за пределами всех объектов. В данном исследовании предлагается использовать одну из наиболее популярных моделеи маркировок - CRF (Conditional Random Fields), подробно рассмотренную в статье "An Introduction to Conditional Random Fields".

2 Постановка задачи

Цель работы состоит в извлечении онтологических объектов из текста с последующим исследованием зависимости качества их распознавания от глубины гипонимии. Для этого используются тексты на англиискомязыке с последующим извлечением именованных сущностеи. В задаче даны предложения в виде последовательности токенов $w = = (w1, w2, ..., w_n)$, и мы должны вывести последовательность тэгов $y = (y1, y2, ..., y_n)$. Для этого используется BIO формат разметки и каждый онтологический объект получает тэг, соответствующий данному уровню гипонимии. Для каждои сущности строится цепочка соответствующих гиперонимов на основе данных из системы WordNet. Наборы данных отличаются в зависимости от заданнои глубины гипонимии. Задача решается с помощью стандартных моделеи NER.

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

В качестве результатов данной работы можно отметить следующее:

- ✓ Было предложено решать задачу выделения онтологических объектов методами, используемыми в задаче поиска именованных сущностей.
- ✓ Получена пробная разметка, дающая датасеты для проведения экспериментов, проверяющих зависимость качества используемых методов от параметра глубины гипонимии.
- ✓ В результате экспериментов было получено, что с увеличением уровня гипонимии качество распознавания онтологических объектов падает.

Мы надеемся, что результаты данной работы могут быть использованы для решения проблемы поиска онтологических объектов и аналогичных задач..

6 Обзор литературы

Существует множество моделей NER, в том числе методов основанных на их комбинациях. Например, рассмотренный в статье “CollaboNet: collaboration of deep neural networks for biomedical named entity recognition”, метод для распознавания биомедицинских именованных сущностей -
CollaboNet, основан на комбинации нескольких моделей NER: Recurrent Neu- ral Network (RNN), bidirectional Long Short-Term Memory Network s(BiLSTM) и Conditional Random Field (CRF). Его можно использовать для уменьшения количества ложных определений и ошибочно классифицированных объектов, включая многозначные слова.
В CollaboNet модели, обученные на разных наборах данных, соединяются друг с другом. В работе были использованы такие наборы данных : BC2GM [33], BC4CHEMD [34], BC5CDR [35, 36, 37, 38], JNLPBA [22], NCBI [21]. Минус модели заключается в том, что нехватка данных может отрицательно влиять на точность результата.
Другим примером модели является, рассмотренный в статье “ Efficient Dependency-Guided Named Entity Recognition”, - NER с функциями зависимости. Его особенность состоит в том, что отношения между объектами представлены деревьями зависимостей. Для работы NER используется semi-CRF. Тема такой структуры уже была рассмотрена в предыдущем исследовании авторов и показала что данные функции работают достаточно эффективно. Использованные наборы данных : Broadcast News section (OntoNotes dataset release 5.0), включающие 6 секций: ABC, CNN, MNB, NBC, PRI and VOA.
CRF- фреймворк для построения вероятностных моделей для сегментации и создания меток sequence data. Используется для улучшения качества(точности) меток. Подробнее вопрос раскрывается в работе “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. CRF могут быть обучены с использованием exponential loss

objective function применяемой AdaBoost алгоритмом. Другим привлекательным аспектом фреймворка является то, что можно реализовать эффективный выбор функций и алгоритмы индукции для них. Лингвистические особенности слова, такие как наклонения, падежи или другие грамматические случаи, могут использоваться в NER для улучшения качества работы для языков, богатых морфологией. Однако этот метод требует использования внешних инструментов морфологической неоднозначности , которые трудно получить или которые вообще не существуют для многих языков. В работе “Improving Named Entity Recognition by Jointly Learning to Disambiguate Morphological Tags” предлагается модель, которая облегчает потребность в таких неоднозначных элементах, совместно изучая NER и MD(Morphological Disambiguation)-метки на языках, для которых можно предоставить список возможных морфологических анализов. Эксперименты проводились с использованием трех разных архитектур моделей.

Первая - Bi-LSTM слой , который снабжается представлением слов, как в основных моделях, NER и MD.Вторая модель также вычисляет отдельные потери для каждой задачи и суммирует их для получения единой потери для оптимизации. Последняя архитектура использует три Bi-LSTM(bidirectional long short-term memory) слоя вместо одного. Наиболее оптимальным оказался первый вариант.

Модель с доминирующей нейронной архитектурой, состоящей из двунаправленной рекуррентной нейронной сети рассмотрена в “Deformable Stacked Structure for Named Entity Recognition”. На основе данного алгоритма было проведено три эксперимента: деформируемая сложная структура (deformable stacked structure) между слоями BiLSTM, деформируемая сложная структура между слоем кодирования (BiLSTM) и слоем декодера (CRF), и последний эксперимент проводился на основе первых двух моделей.

Литература

[1] An Introduction to ConditionalRandom Fields
https://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf

[2] CollaboNet: collaboration of deep neural networks for biomedical named entity recognition
https://arxiv.org/pdf/1809.07950v1.pdf

[3] Deformable Stacked Structure for Named Entity Recognition
https://arxiv.org/pdf/1809.08730v2.pdf

[4] Efficient Dependency-Guided Named Entity Recognition
https://arxiv.org/pdf/1810.08436.pdf

[5] Improving Named Entity Recognition by Jointly Learning to Disambiguate Morphological Tags
https://arxiv.org/pdf/1807.06683v1.pdf

[6] Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition
https://arxiv.org/pdf/1709.09686.pdf

[7] A Byte-sized Approach to Named Entity Recognition
https://arxiv.org/pdf/1809.08386v1.pdf

[8] On the Strength of Character Language Models for Multilingual Named Entity Recognition
https://arxiv.org/pdf/1808.08450v1.pdf

[9] Named entity recognition: a maximum entropy approach using global information
https://dl.acm.org/citation.cfm?id=1072253

[10] Named Entity Recognition with Bidirectional LSTM-CNNs
https://arxiv.org/pdf/1511.08308.pdf

[11] Named entity recognition using conditional random fields with non-local relational constraints
https://arxiv.org/pdf/1310.1964v1.pdf