# Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии.*

***Кузнецов М. Д., Дочкина В., Резяпкин В., Русскин А., Ярмошик Д.***
kuznetsov.md@phystech.edu

Данная работа посвящена изучению онтологических объектов и их гипонимий. Проведен сбор датасета, состоящий из текстов на английском языке, анализ и разметка данных текстов по нескольким уровням вложенности. Также (будет) проведена серия экспериментов для определения зависимости качества распознавания объектов от уровня вложенности. Обученные модели, полученые в результате данных экспериментов, могут использоваться в получении дополнительных признаков для решеня различных задач NLP. Исследованная методика разметки онтологических объектов в свою очередь может быть использована для автоматической разметки других текстов.

**Ключевые слова**: *гипонимия, онтология, онтологические объекты.*

## 1. Введение

Распознование именнованных сущностей (NER - Named entity recognition) является важной задачей NLP (Natural Language Processing — обработка естественного языка), которая заключается в автоматическом распознавании объектов в тексте и классификация по определенным типам сущностей, такие как люди, организации, геополитические субъекты, места, события и т. д. К примеру, в предложении “Jim bought 300 shares of Acme Corp. in 2006.” модель NER разметки определит слово “Jim” как “Person”, “Acme Corp.” как “Organization”, а “2006” как “Time”, остальные же слова оставит “Out” (out-of-tag). NER является фундаментальным компонентом многих методов извлечения информации, включая извлечение связей, “ответов на вопросы” и др. В данной работе рассматривается немного отличная от NER задача. Рассматриваются не именнованные сущности, а так называемые онтологические объекты. Онтологическим объектом является любое слово, обозначающее какую-либо сущность, явление, предмет и др. Один из аспектов “связывания” объектов есть отношения обобщений, или гипонимия – иерархическая организация элементов, основанная на родо-видовых отношениях. Выделяют два определения: гипоним – понятие, выражающее частную сущность по отношению к другому, более общему понятию; и гипероним – понятие, выражающее общее, родовое понятие, название класса (множества) предметов (свойств, признаков). К примеру, слово “seat” является гиперонимом слову “chair” и гипонимом слова “furniture” . Каждая сущность может быть обобщена до более высокого уровня вложенности понятий. Поставленная задача заключалась в том, чтобы создать некоторое количество датасетов, состоящих из одинаковых текстов, но имеющих разную глубину разметки по гипонимиям, а также проанализировать связи между уровнями вложенности обобщений для одинаковых онтологических объектов. Результаты проведенного анализа можно использовать для повышения качества выполнении задач NLP, а также разработки новых возможностей для уже существующих моделей. Данная модель (в будущем и возможно) способна анализировать контекст слова и давать именно ту ветвь вложенности обобщений, которая имеется в виду по смыслу в тексте или предложении. Для проведения эксперимента и исследования в качестве изначальных датасетов

---

Научный руководитель: Стрижов В. В. Задачу поставил: Бурцев М. С. Консультант: Баймурзина Д. Р.

взяты (еще под вопросом, какие именно) тексты, для разметки взят английский словарь WordNet, включающий информацию о гипонимах и гиперонимах.

## 2. Обзор литературы

В решении задачи NER применялось множество различных методов и моделей. В сентябре 2018 года были представлены две модели для решения проблемы необходимости большого количества вручную размеченных данных, включающие в себя нейронную модель AutoNER с "Tie or Break scheme" [3]. В USC были проведы исследования, направленные на разметку не только слов, но и частей слов [4]. Также, результаты решения NER задач были использованы для определения эмоциального окраса сообщений в социальных сетях [5]. В том числе, были исследования, направленные на увеличение точности, а также многоязычности посредством использования Википедии [6]. Задача, поставленная в данной работе, несколько отличается от задачи NER, но модели и наработки можно использовать и/или модернизировать для решения.

## 3. Постановка задачи

Необходимо рассмотреть различные сущности, их гиперонимы; собрать и обработать подходящие друг другу для различных сущностей уровни гипонимии для онтологических объектов, на которые впоследствие эти объекты будут распределены, создать словари на базе WordNet. Следующий шаг – анализ выбранных уровней на оптимальность . Необходимо также провести обработку датасетов, путем разметки выбранных текстов по разным уровням гипонимии. Завершающим этапом исследования является проведение некоторых экспериментов и анализ данных датасетов, выбор оптимальной глубины для различных задач.

## 4. Описание основных методов

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

### 4.1. Рекуррентные нейронные сети

Для учета контекста в тексте используются рекуррентные нейронные 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 – нейросетевой рекуррентный блок, состоящий из элементов: Основной слой (как и в обычной RNN), три сигмоидальных слоя-фильтра, Ячейка памяти. Формулы для этих компонент:

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

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

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

$$c_t = f_t \circ c_{t-1} + i_t \circ c_n, \tag{4}$$

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

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

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:
$\sigma_g$ : на основе сигмоиды.
$\sigma_c$ : на основе гиперболического тангенса.
$\sigma_h$: 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

## 4.2. 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 прошлый контекст, двунаправленная учитывает и будущий. 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.

## 4.3. 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.

## 4.4. 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.

## 5. 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** |

## 6. Заключение

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

## Литература

[1] *Воронцов К. В.* LaTeX $2_\varepsilon$ в примерах. 2006. `http://www.ccas.ru/voron/latex.html`.

[2] *Львовский С. М.* Набор и вёрстка в пакете LaTeX. 3-е издание. Москва: МЦНМО, 2003. 448 с.

[3] *Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, Jiawei Han* Learning Named Entity Tagger using Domain-Specific Dictionary

[4] *Emily Sheng, Prem Natarajan.*A Byte-sized Approach to Named Entity Recognition

[5] *Dilek Küçük* Joint Named Entity Recognition and Stance Detection in Tweets

[6] *Jian Ni, Radu Florian* Improving Multilingual Named Entity Recognition with Wikipedia Entity Type Mapping

[7] *Bill Yuchen Lin, Wei Lu* Neural Adaptation Layers for Cross-domain Named Entity Recognition