

# Исследование зависимости качества распознавания онтологических объектов от глубины гипонимии

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

МФТИ

10 декабря 2018

# План

- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение





- Гипонимия - вид системных отношений в лексике: связь слов по линии «общее и частное». Пример.
- Целью исследования является выделение объекта в тексте и последующего определения его к конкретному типу, то есть сопоставление гипонима его гиперониму
- Информация об онтологических объектах оказывается полезной в задачах, связанных с обработкой естественного языка, включая ответы на вопросы, заданные пользователем и извлечении зависимостей между объектами

# План

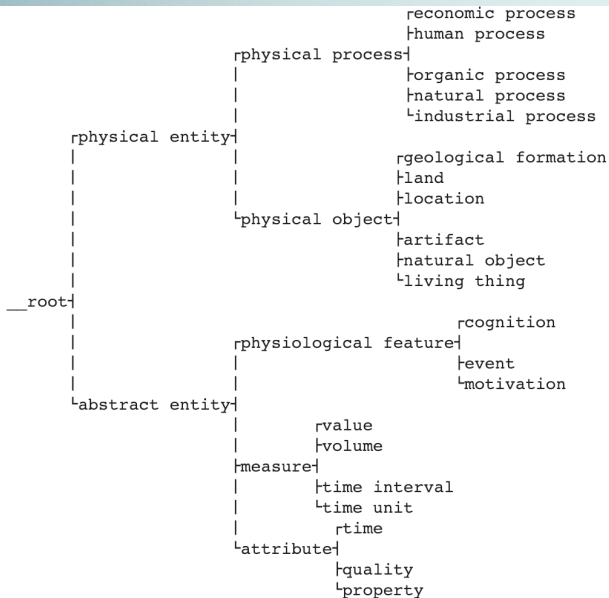
- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение





- 1 Задача NER
- 2 Пример произвольно размеченного текста:  
"[ORGANIZATION MIPT], is a Russian university,  
located in [LOCATION Dolgoprudny]"
- 3 Дано: предложения в виде последовательности токенов  $w = (w_1, w_2, \dots, w_n)$ , и мы должны вывести последовательность тэгов  $y = (y_1, y_2, \dots, y_n)$
- 4 Используется BIO формат разметки и каждый онтологический объект получает тэг, соответствующий данному уровню гипонимии. Для каждой сущности строится цепочка соответствующих гиперонимов на основе данных из системы WordNet.

Figure: Дерево гипонимии





# Dictionary

geological formation

land

location

living thing

natural object

artifact

economic process

human process

industrial process

natural process

organic process

cognition

motivation

event

time

property



# План

- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение







- ① “Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition”
- ② "An Introduction to Conditional Random Fields"
- ③ "Improving Named Entity Recognition by Jointly Learning to Disambiguate Morphological Tags"
- ④ "Collaboration of deep neural networks for biomedical named entity recognition"

# План

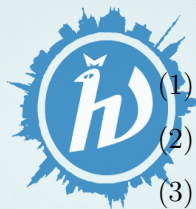
- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение





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

Для учета контекста в тексте используются рекуррентные нейронные сети (RNN). 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:



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

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

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

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

$$h_t = o_t \circ g(c_t), \quad (5)$$

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

Переменные:

$x_t$  - входной вектор,  $h_t$  — выходной вектор

$c_t$  — вектор состояний,  $W, U, b$  — матрицы параметров и вектор

$i_t$  — 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.

Сигмоида :

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

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

Гиперболический тангенс

$$f(x) = \tanh(x) \quad (8)$$

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





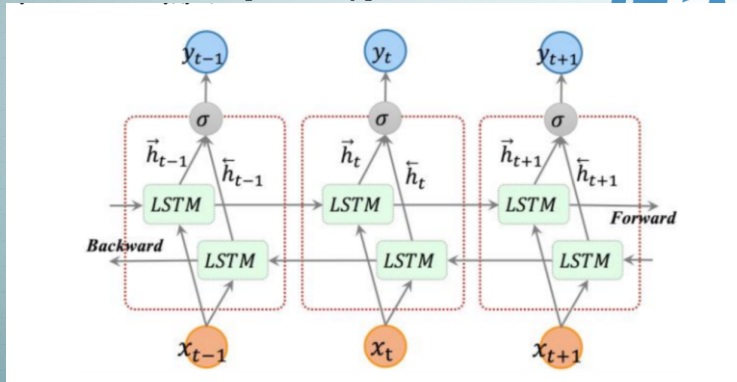


- 1 Bi-LSTM разработаны для кодирования каждого элемента в последовательности с учетом левого и правого контекстов, что делает его одним из лучших вариантов для задачи NER
- 2 Обычная LSTM учитывает только прошлый контекст, двунаправленная учитывает и будущий
- 3 Двунаправленные рекуррентные сети полезны во многих задачах NLP, таких как: машинный перевод, ответ на вопросы, и особенно для решения проблемы NER



- 1 Обычная LSTM учитывает только прошлый контекст, двунаправленная учитывает и будущий:

## Архитектура Bi-LSTM





- 1 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  $\vec{y} = y_0, y_1, \dots, y_n$  тегов с учетом предложения  $\vec{x} = x_0, x_1, \dots, x_n$ . Для этого вычисляется условная вероятность:

$$p(\vec{y} \mid \vec{x}) = \frac{e^{\text{Score}(\vec{x}, \vec{y})}}{\sum_{y'} e^{\text{Score}(\vec{x}, y')}} \quad (9)$$

где Score рассчитывается по формуле :

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

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



- 1 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 :

$$F_1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (11)$$

Recall - полнота,

Precision - точности

# План



- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение

# Эксперимент и его результаты



- 1 Система wordnet для получения информации о деревьях гипонимов
- 2 датасеты Web of Science, состоящие из текстов на английском языке
- 3 для каждого уровня гипонимии создан отдельный датасет с разметкой формата IOB(Inside–outside–beginning)
- 4 модель Bi-List+CRF для обучения

## Результаты:

Table:  $F_1$  мера для разных уровней гипонимии  
и для первого датасета

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



Table:  $F_1$  мера для разных уровней гипонимии для второго  
датасета

Первый уровень:	90.77
Второй уровень:	81.45
Третий уровень:	68.81



# План

- 1 Цель исследования
- 2 Постановка задачи
- 3 Литература
- 4 Теоретическая часть
  - Элементы LSTM
  - Функции активации
  - Bi-LSTM
  - Архитектура Bi-LSTM
  - CRF
  - F-мера
- 5 Эксперимент и его результаты
  - Эксперимент и его результаты
- 6 Заключение





Результаты проведенного эксперимента показывают, что с увеличением уровня гипонимии качество распознавания онтологических объектов падает. Однако, для датасета с достаточно большим количеством обучающих данных, второй уровень гипонимии дает сравнительно неплохой результат.