Задача поиска символов в текстах*

***Северилов П. А.*[1], *Лемтюжникова Д. В.*[1], *Апишев М. А.*[2]**
severilov.pa@phystech.edu, daratigra@icloud.com, great-mel@yandex.ru
[1]Московский физико-технический институт (МФТИ);
[2]Московский государственный университет имени М.В.Ломоносова (МГУ им. М.В.Ломоносова)

В работе рассматривается задача поиска символов в тексте. Определение в тексте средств выразительности таких, как метафоры, аллегории и пр. у экспертов происходит в ручном режиме, и процесс никак не автоматизирован. В простейшем случае эта задача сводится к проблеме Sequence Labeling на размеченной выборке. Сравниваются работы современных алгоритмов решения Sequence Labeling и определяется применимость данных методов к задаче.

Ключевые слова: *распознавание символов, LSTM нейронные сети, sequence labeling, скрытые марковские модели.*

1 Введение

Современные модели для обработки текстов не справляются с главной особенностью языка — неоднозначностью смысла высказывания. Текст воспринимается ими буквально, и различные средства художественной выразительности, в частности, символы, метафоры, аллегории и др. не интерпретируются очевидным образом. Так, например, выражение «золотые руки» вероятнее всего будет понято моделью, как «руки из золота» вместо верного «умения очень хорошо делать что-либо». Автоматизация поиска выражений с неоднозначным смыслом продвинет механизм обработки текстов на другой уровень абстракции.

Задача поиска символов в тексте может быть сведена к sequence labeling – более общей задаче, широко распространенной в NLP. Как правило, рассматриваются три конкретных типа sequence labeling: тегирование частей речи (part-of-speech tagging), распознавание именованных сущностей (named entity recognition) и синтаксический анализ (shallow parsing). Данные методы могут быть применены к текущей задаче.

Традиционно задачи sequence labeling решаются с использованием линейных статистических моделей, например: скрытые марковские модели (HMM), марковские случайные поля (CRFs). Реализация решений происходит с помощью различных архитектур нейронных сетей. В данной статье сравниваются результаты работ нескольких state-of-the-art архитектур для sequence labeling применительно к задаче поиска символов.

Архитектуры таких нейронных сетей главным образом базируются на рекуррентных сетях, а именно Bidirectional LSTM с использованием CRF. Рассматривается три подхода. Первый – это классическая реализация BiLSTM. Затем предложена следующая модификация: на вход сети будут подаваться не векторные представления слов в целом, а каждый символ по отдельности. И третий подход – основан на предыдущей модели, но с интеграцией механизма внимания (attention) в архитектуру.

Чтобы определить скрытый смысл высказывания, для начала необходимо ответить на вопрос, есть ли он вообще в нём. Поэтому решается задача классификации символ/не символ. Тестирование проводится на тексте романа М. А. Булгакова "Мастер и Маргарита"и стихотворных текстах поэтов серебряного века. Основная сложность заключается в получении достаточного объёма обучающих данных, то есть требуется по имеющейся

* Задачу поставил: Лемтюжникова Д. В., Консультант: Апишев М. А.

небольшой экспертной разметке получить выборку большего размера. Наличие разметки позволит провести эксперименты с подбором оптимальной модели и в целом определить применимость данных методов для решения задачи поиска символов.

2 Постановка задачи

Рассматривается два подхода к задаче: с точки зрения классификации и с точки зрения sequence labeling.

1. **Sequence labeling**: Дано предложение $\mathbf{X}$, разделенное на части (слова): $\{x_1, x_2, \cdots, x_n\}$. Требуется построить последовательность двоичных меток (labels) $\{l_1, l_2, \cdots, l_n\}$, которые идентифицируют наличие неоднозначности/символа в каждом слове x_i
2. **Классификация**: Аналогично дано предложение $\mathbf{X}$, разделенное на части (слова): $\{x_1, x_2, \cdots, x_n\}$. Требуется для целевой переменной i предсказать отношение x_i к классу символ или не символ, соответственно 1 и 0.

В целом, Sequence labeling является обобщением классификации в данном случае, поэтому в общем задачу можно описать так: $\mathbf{X}$ – множество слов предложения, $\mathbf{Y}$ – множество ответов (отношение к классу 1 или классу 0). Требуется построить алгоритм $a : \mathbf{X} \to \mathbf{Y}$ способный классифицировать произвольный объект $x_i \in X$

3 Базовый алгоритм

3.1 Bidirectional LSTM for sequence labeling

Для начала разберем базовую модель: Bidirectional LSTM. На рисунке 1 изображена общая схема нейронной сети для sequence labeling. Модель получает на вход последовательность слов $(w_1, w_2, \cdots, w_\mathrm{T})$ и предсказывает для каждого из них соответствующую ему метку – символ/не символ. Для начала слова переводятся в векторное пространство (например чере word2vec), в результате чего получается последовательность векторов из этого пространства $(x_1, x_2, \cdots, x_\mathrm{T})$. Далее, эти векторные представления подаются на вход двум LSTM компонентам (Hochreiter, Schmidhuber, 1997), двигаясь по тексту в различных направлениях, таким образом создавая представления для конкретного контекста. Соответствующие прямые и обратные представления конкатенируются для каждого положения слова:

$$\overrightarrow{h_t} = \mathrm{LSTM}(x_\mathrm{t}, \overrightarrow{h_{t-1}}) \qquad \overleftarrow{h_t} = \mathrm{LSTM}(x_\mathrm{t}, \overleftarrow{h_{t+1}}) \qquad h_\mathrm{t} = [\overrightarrow{h_\mathrm{t}}; \overleftarrow{h_\mathrm{t}}]$$

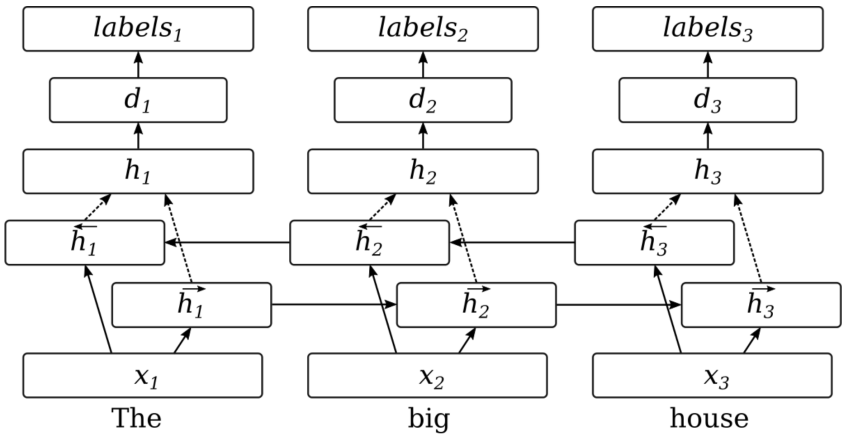

Рис. 1 Схема базовой модели BiLSTM

Затем, добавляется скрытый слой нелинейности:

$$d_{\mathrm{t}} = tanh(W_{\mathrm{d}} h_{\mathrm{t}}),$$

где $\mathbf{W}_{\mathrm{d}}$ – весовая матрица между слоями.

В конце для создания самих меток используется либо softmax либо Conditional Random Fields (разновидность метода Марковских случайных полей) в зависимости от выбранной постановки задачи. Функция softmax рассчитывает нормированное распределение вероятностей по всем возможным меткам для каждого слова:

$$\mathbb{P}(y_t = k|d_t) = \frac{e^{W_k d_{\mathrm{t}}}}{\sum_{\tilde{k} \in K} e^{W_{\tilde{k}} d_{\mathrm{t}}}},$$

где $\mathbb{P}(y_t = k|d_t)$ – вероятность того, что метка t-ого слова y_t будет k (K – множество всевозможных меток), $\mathbf{W}_k$ – k-ая строка весовой матрицы $\mathbf{W}$.

Для оптимизации модели используется минимизация отрицательного логарифма вероятности верной метки:

$$\mathcal{L} = -\sum_{t=1}^{T} \log(\mathbb{P}(y_t|d_t))$$

Т.е. решается данная задача:

$$\mathbf{W}^* = \operatorname*{argmin}_W(\mathcal{L}(\mathbf{W}))$$

3.2 Усовершенствования базового алгоритма

4 Результаты работы алгоритма

Самый важный график

5 Заключение

Литература

[1] *Marek Rei, Gamal K.O. Crichton, Sampo Pyysalo N.* Attending to Characters in Neural Sequence Labeling Models // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers , 2016, C16-1030, Pp. 309–318.

[2] *Adnan Akhundov, Dietrich Trautmann, Georg Groh N.* Sequence Labeling: A Practical Approach // CoRR , vol. abs/1808.03926, 2018.

[3] *Zachary Chase Lipton, John Berkowitz N.* A Critical Review of Recurrent Neural Networks for Sequence Learning // CoRR , vol. abs/1506.00019, 2015.

[4] *Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia N.* Attention is all you need // Advances in Neural Information Processing Systems 30, 2017, Pp. 5998–6008.

[5] *Adnan Akhundov, Dietrich Trautmann, Georg Groh N.* LONG SHORT-TERM MEMORY // Journal Neural Computation archive Volume 9 Issue 8 , 1997, Pp. 1735–1780