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 был автоматизирован. В простейшем случае эта задача сводится к проблеме Sequence Labeling на размеченной выборке. Мы сравниваем работы нескольких современных алгоритмов решения Sequence Labeling и определяем применимость данных методов к нашей задаче. Также предложена метрика качества классификатора для определения символ/не символ.

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. никак не интерпретируются в правильном смысле. 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 обработку текстов на более глубокий уровень понимания текста у моделей.

Данная 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 вполне подходят к нашей 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 мы сравниваем результаты работы 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 конкретнее Bidirectional LSTM с использованием CRF. Рассматриваем три подхода. Первый – это классическая реализация BiLSTM. Затем аналогичная архитектура с изменением в том, что на вход сети будут подаваться не векторные представления слов в целом, а каждый символ по отдельности. И последняя – основанная на идее предыдущей модели, но с добавлением механизма внимания (attention) в архитектуру (Transformer).

Для того, чтобы определить скрытый смысл высказывания, для начала нужно научиться понимать, а есть ли он вообще в конкретном выражении. Поэтому мы решаем задачу классификации символ/не символ. Тестирование проводится на тексте романа

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

М.А.Булгакова "Мастер и Маргарита". Основная сложность заключается в получении достаточного объёма обучающих данных, то есть требуется по имеющейся 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 Название раздела

smth

2.1 Название параграфа.

another smth

2.2 Теоретическая часть

3 Заключение

final