Задача поиска символов в текстах

Северилов Павел Андреевич

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 674, весна 2019

Цель исследования

Проблема

Современные модели для обработки текстов воспринимают высказывания буквально, и различные средства художественнои выразительности, в частности, символы, метафоры, аллегории и др. не интерпретируются ими верным образом.

Цель работы

Получить оптимальную модель для определения неоднозначности в высказываниях.

Постановка задачи

Sequence labeling

Дано предложение $\mathbf{X}$, разделенное на 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

Классификация

Аналогично дано предложение $\mathbf{X}$, разделенное на части: $\{x_1, x_2, \cdots, x_n\}$. Требуется для целевой переменной i предсказать отношение x_i к классу символ или не символ, соответственно 1 и 0.

Базовый алгоритм

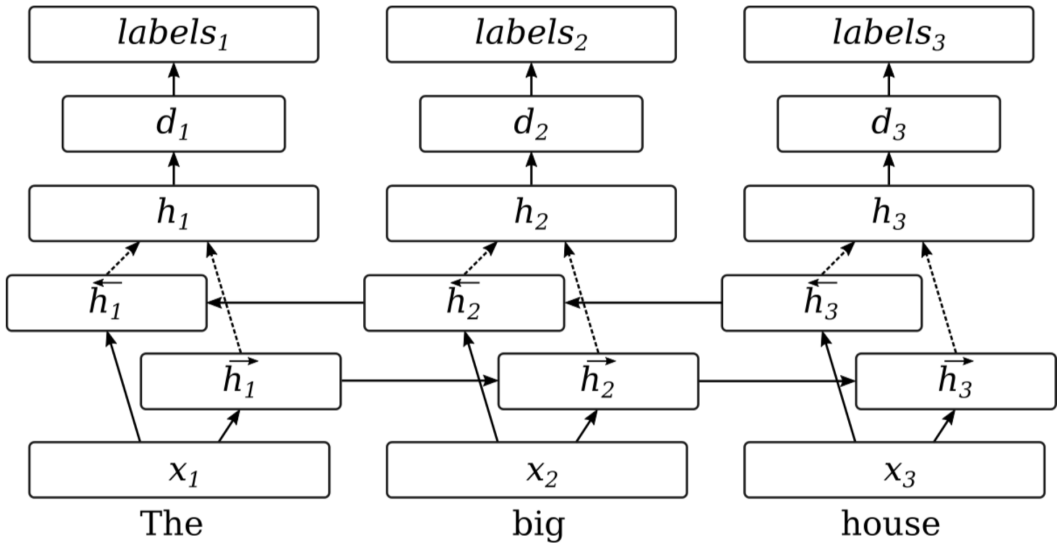


Рис.: Схема базовой модели BiLSTM

Представления в LSTM-сети

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

Скрытый слой нелинейности: $d_t = tanh(W_d h_t)$, где $\mathbf{W_d}$ – весовая матрица между слоями.

Итоговая задача оптимизации

Нормированное распределение вероятностей по всем возможным меткам для каждого слова (softmax):

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

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 функции

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

Т.е. решается данная задача:

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

Вычислительный эксперимент

Модели

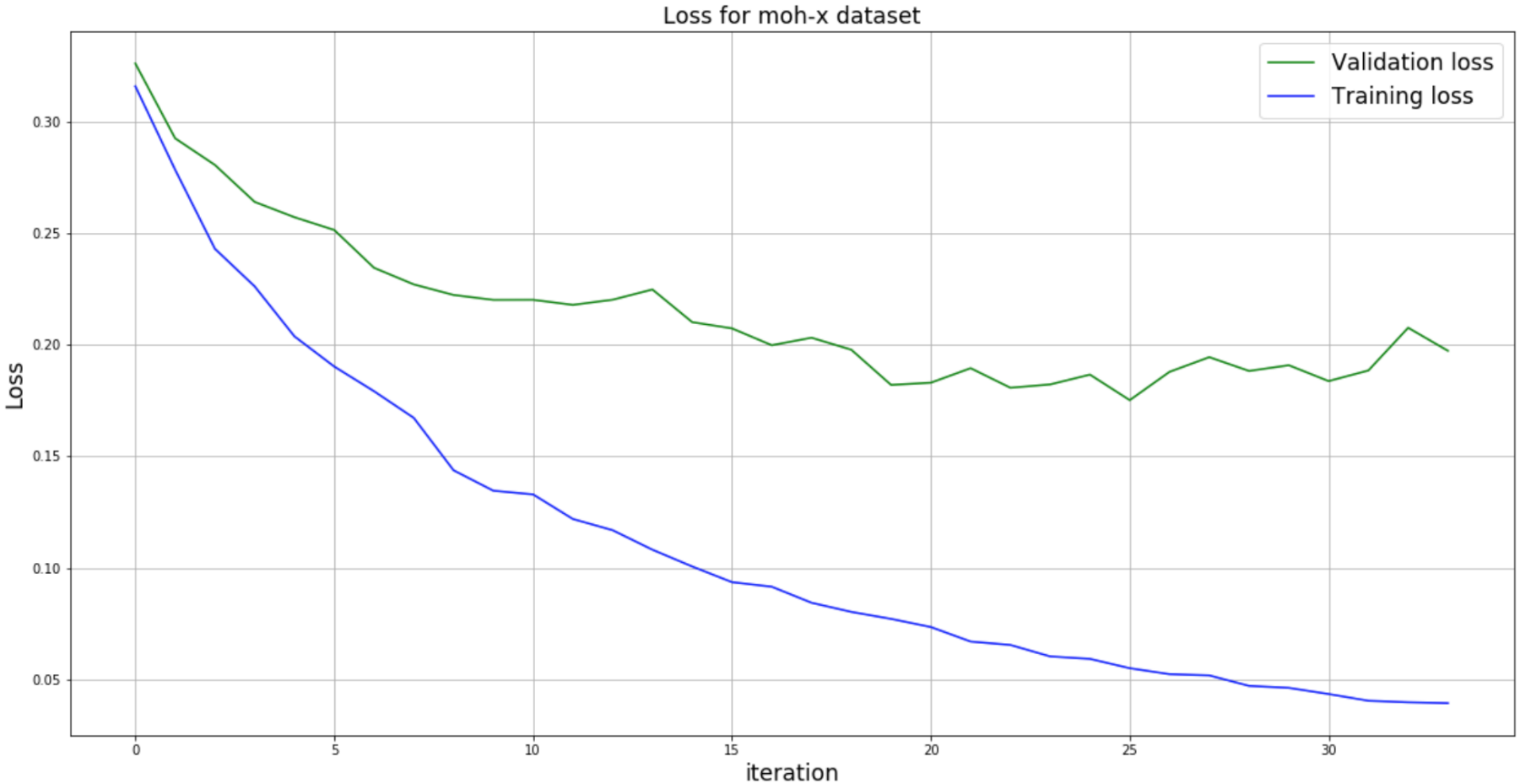
Предлагается сравнить следующие модификации базового алгоритма:

1. Базовая BiLSTM нейронная сеть
2. BiLSTM нейронная сеть с CRF (Conditional random field)
3. BiLSTM нейронная сеть с Attention

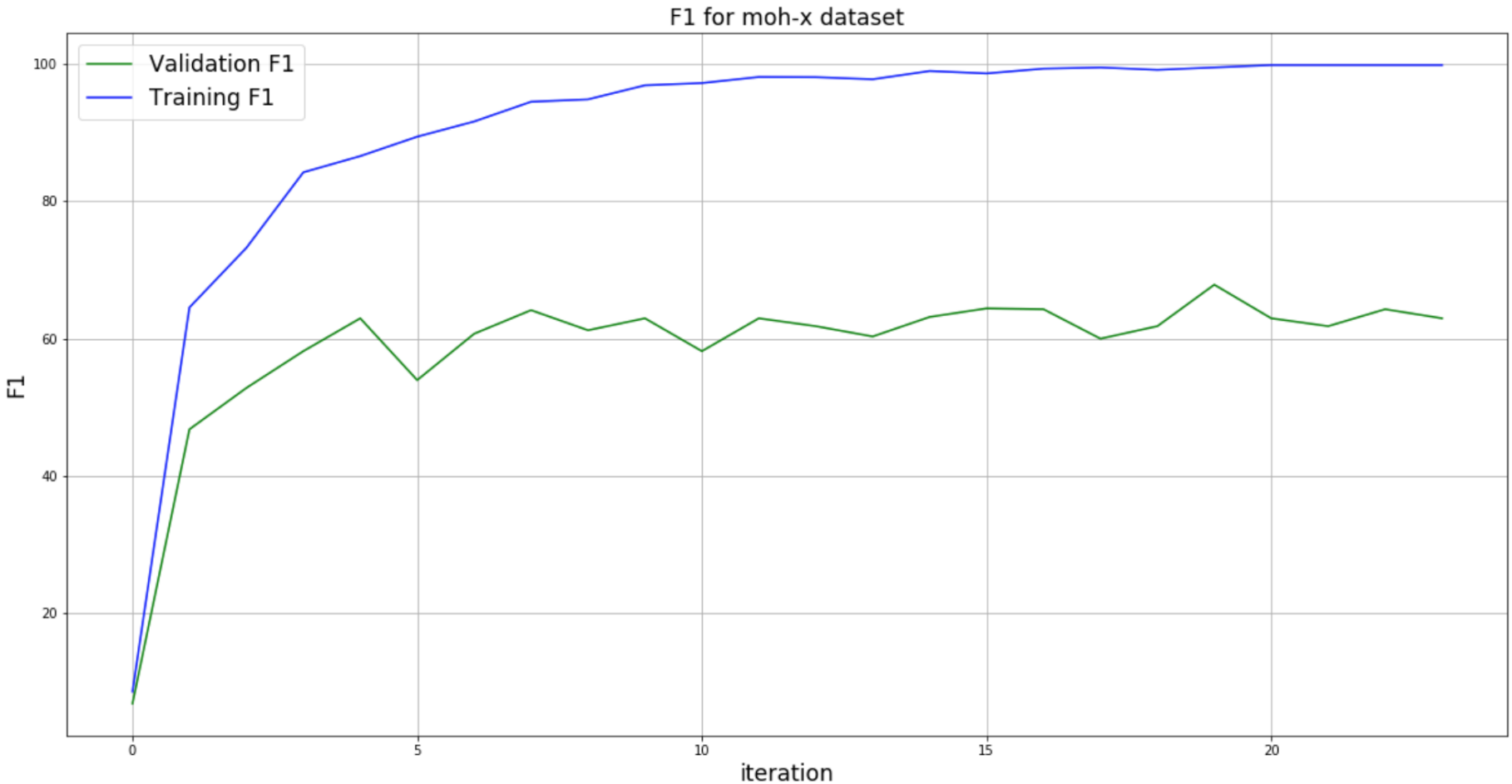
Данные

1. МОН датасет с метафорами (eng)
2. VU Amsterdam Metaphor Corpus (eng)
3. Собственная разметка русских текстов (НОВИНКА)

Результаты эксперимента: Loss график



Результаты эксперимента: F1 мера



Результаты эксперимента: Качество

Качество алгоритма после 30 эпох обучения:

- Precision on MOH = **64**.14203612479474
- Recall on MOH = **67**.85714285714286
- F1 on MOH = **65**.8939014202172
- Accuracy on MOH = **69**.27083333333333

Заключение

- Алгоритм sequence labeling хорошо подходит для поиска символов в тексте
- Сравнены результаты работ нескольких моделей: наилучшая – ?
- Главная проблема эксперимента – мало данных
- Качество заметно улучшится при увеличении выборки
- Предложенные модели могут быть применены для определения не только каких-то конкретных неоднозначностей в тексте, а в целом для всех видов символов
- Для русскоязычных текстов данная задача никак до этого не решалась