Разработка и реализация нейросетевой системы для извлечения именованных сущностей генов и мутаций из медицинских текстов

Сотников А.Д.

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.т.н. С. К. Дулин
Научный консультант д.т.н. В. Ф. Хорошевский

Москва,
2020 г.

О задаче

Мотивация

Существующие модели не способны одновременно находить в текстах сущности и генов, и мутаций. Кроме того, они не используют в полной мере контекстную информацию.

Поставленная задача

На вход подается корпус биомедицинских текстов, разбитый на предложения. Требуется найти в нем именованные сущности генов и мутаций, решив задачу разметки последовательности слов.

Предлагается

Реализовать нейронную сеть, которая с помощью контекстной информации каждого слова способна извлекать из биомедицинских текстов рассматриваемые именованные сущности.

Сложности

Именованные сущности генов и мутаций могут быть записаны в разных форматах:

- **Standard** – c.925delA; g.3912G>C; rs206437.
- **Semi-standard** – 3992-9g–>a mutation; codon 92, TAC–>TAT.
- **Natural language** – deletion of 10 and 8 residues from the N- and C-terminals.

Серьезные трудности возникают, когда сущность специфицируется на естественном языке.

Существующие методы

Работы по Gene Mention

- Vocabulary-based *(BLAST 2000)*
- Rule-based *(ProMiner 2005)*
- ML-based *(GNormPlus 2015)*
- DL-based *(CollaboNet 2018, BioBERT 2019)*

Работы по Mutation Mention

- Rule-based *(MutationFinder 2007, SETH 2016)*
- Probabilistic-based *(tmVar 2013, nala 2017)*

Постановка задачи

- Дана выборка и множество меток

$$\mathcal{D} = \{\mathrm{x}_i, \mathrm{y}_i\}_{i=1}^{N}, \quad \mathcal{K} = \{k_j\}_{j=1}^{5}$$

где $x_i \in \mathbb{R}^{n_i}$ – последовательность слов в предложении длины n_i, $y_i \in \mathbb{R}^{n_i}$ – соответствующая им последовательность меток, а $k_j \in [\mathrm{O},\ \mathrm{B\text{-}GENE},\ \mathrm{I\text{-}GENE},\ \mathrm{B\text{-}MUT},\ \mathrm{I\text{-}MUT}]$.

- Требуется построить модель

$$a : (\mathrm{w}, \mathrm{X}) \to \mathrm{y},$$

$\mathrm{w} \in \mathbb{W}$ – параметры модели, $\mathrm{X} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{n_i} x_j, \mathrm{y} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{n_i} y_j$.

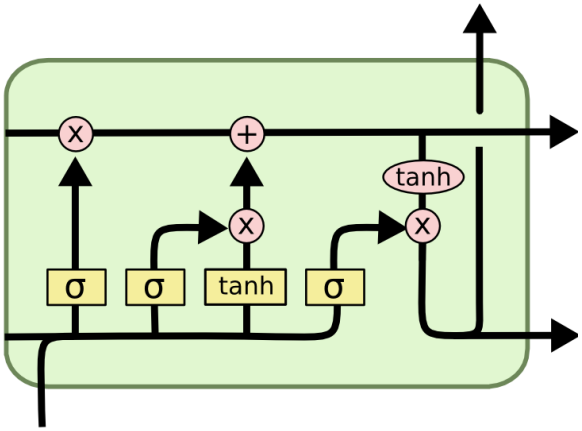
- Функция ошибки

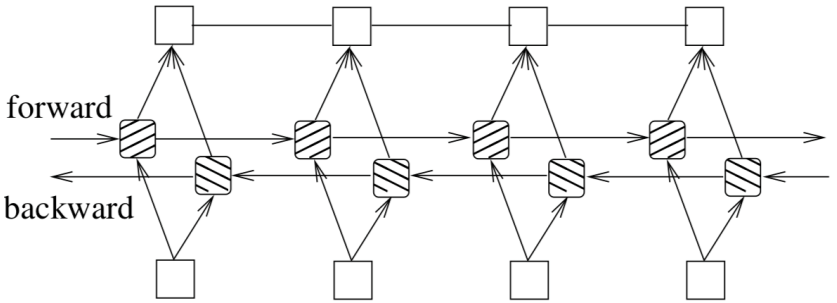
$$\mathcal{L}(\mathrm{y}, \mathrm{X}, \mathrm{w}) = -\sum_{i=1}^{N} \log(\mathbb{P}\{\mathrm{y_{ik}} | \mathrm{x_i}\}).$$

- Решается задача оптимизации:

$$\mathbf{w}^* = \underset{\mathrm{w} \in \mathbb{W}}{\operatorname{argmin}} \big(\mathcal{L}(\mathbf{w})\big)$$

Архитектура BiLSTM





Принцип работы

LSTM помогает сохранять предыдущую относительно текущего слова информацию. При этом, можно параллельно обучить два таких слоя для прямого и обратного предложения, получив скрытое состояние BiLSTM конкатенаций двух состояний LSTM, т.е.

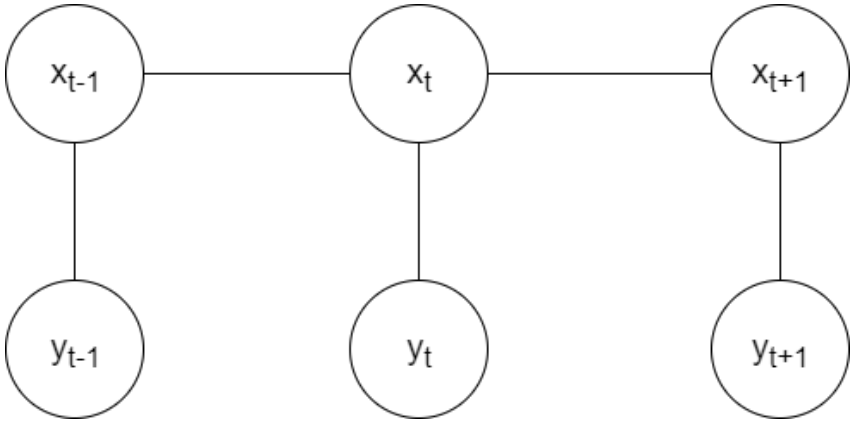
$$\forall x_t \in x \quad h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}].$$

Важно

Такая архитектура позволяет запоминать и учитывать предыдущий и будущий контекст текущего слова.

Условные случайные поля (CRF)

- Имеется $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ – входное предложение, $\mathrm{y} = (y_1, y_2, \ldots, y_n)$ – наблюдаемые метки.



Граф CRF на шаге t

- Оценка последовательности меток y :

$$s(\mathrm{x}, \mathrm{y}) = \sum_{t=1}^{n+1} \left(\mathrm{A}_{y_{t-1}, y_t} + \mathrm{P}_{t, y_t} \right),$$

P_{t,y_t} вычисляет оценку соответствия метки y_t слову x_t (в нашем случае $\mathrm{P}_{t,y_t} = \mathrm{W}_h \cdot \mathrm{h} + \mathrm{b}_h$), A – матрица транзитивности.

- Вероятность последовательности y :

$$\mathbb{P}\{y|x\} = \frac{\prod_n \exp(s\,(x, y))}{\sum_{\tilde{y}\in Y_x} \prod_n \exp(s\,(x, \tilde{y}))},$$

  Y_x – всевозможные последовательности меток в предложении x.

- Окончательно,

$$y^* = \underset{\tilde{y}\in Y_x}{\operatorname{argmax}}\ s\,(x, \tilde{y})$$

Важно

CRF учитывают контекст на уровне предложений. Так, например, они предотвращают ситуации, когда метка I-MUT возникает перед B-MUT в пределах одной сущности или возникает сразу после B-GENE.

Char-level embeddings

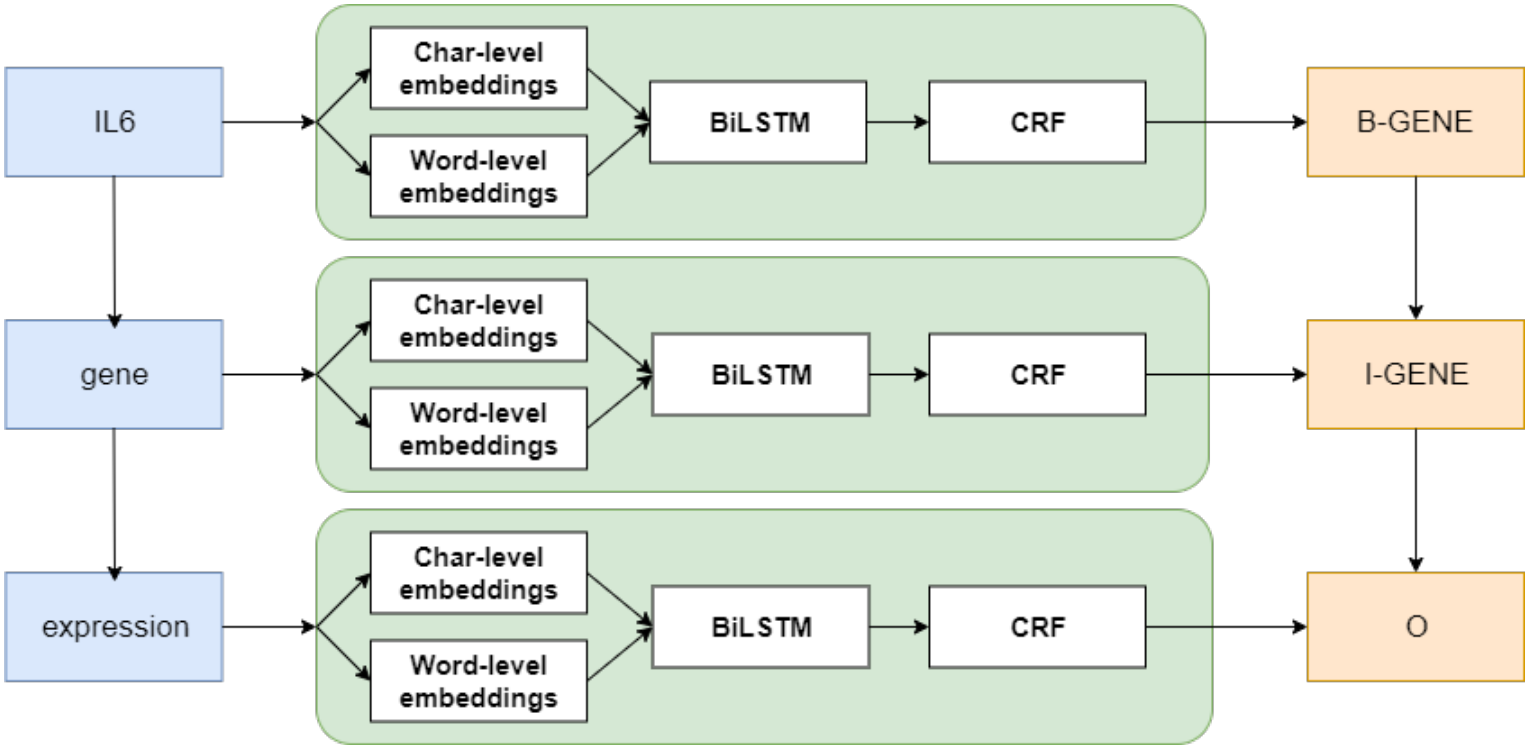
Проблема

Много специфичных слов, не входящих в словарь V $\rightarrow$ возникает проблема OOV (out of vocabulary) слов.

Идея

- Отдельно, с помощью BiLSTM, обучать векторные представления для символов (char-level embeddings), из которых состоит очередное слово.
- Такой метод позволяет модели "приблизительно понять" , какую смысловую нагрузку несёт неизвестное слово.

Модель нейронной сети GMNet



Архитектура нейронной сети GMNet, реализованной в данной работе

Вычислительный эксперимент

Корпус	Кол-во генов	Кол-во мутаций	Кол-во предложений
JNLPBA	10589	-	22562
BC2GM	24583	-	20510
MutationFinder	-	5611	8176
tmVar	-	3702	5956

Наборы данных

Показатели качества

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F\text{-}measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

...C - - > T transition at nucleotide 677 – is one among them...

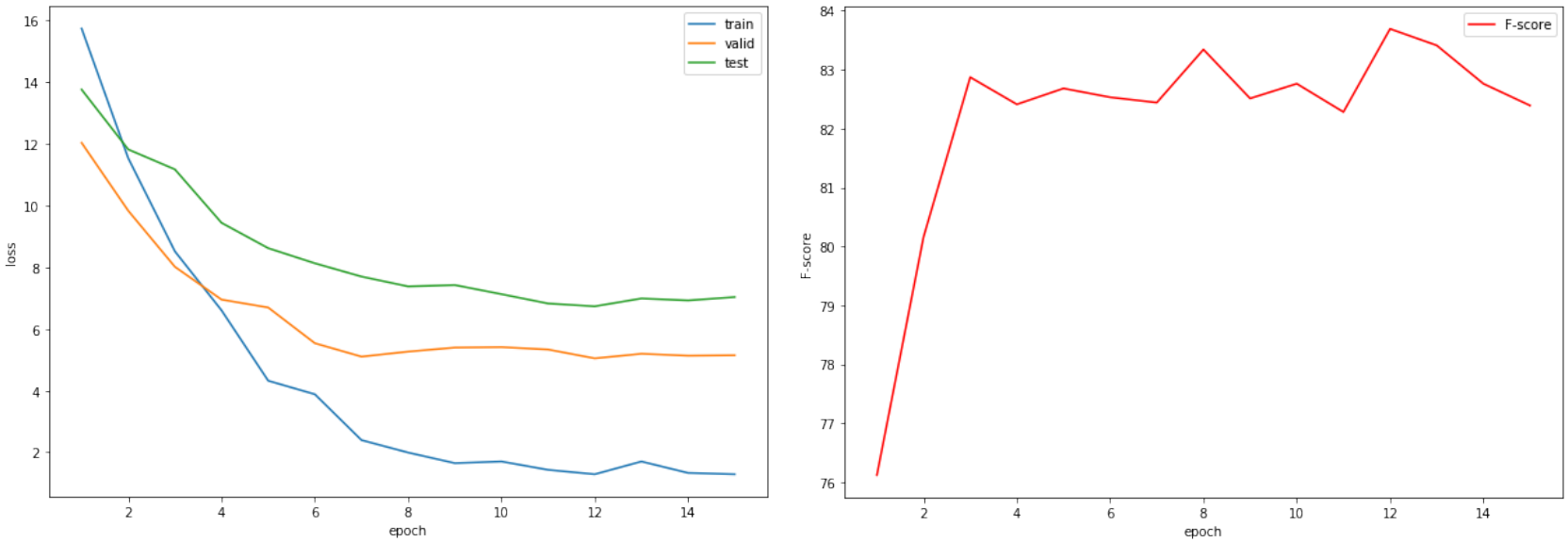
...NF - kappaB DNA – protein binding and ICAM - 0 promoter...

...pRSV-neo and pSV0-neo...

...present study G146A and G146V mutants...

...Human beta0-adrenergic receptor impart...

Результаты обучения на 15 эпохах



Слева: График зависимости функции потерь от числа эпох; Справа: График зависимости F-меры от числа эпох

Пример работы модели

...C - - > T transition at nucleotide 677 – is one among them...

...C - - > T transition at nucleotide 677 – is one among them...

...NF - kappaB DNA – protein binding and ICAM - 0 promoter...

...NF - kappaB DNA – protein binding and ICAM - 0 promoter...

...pRSV-neo and pSV0-neo...

...pRSV-neo and pSV0-neo...

...present study G146A and G146V mutants...

...present study G146A and G146V mutants...

...Human beta0-adrenergic receptor impart...

...Human beta0-adrenergic receptor impart...

Предсказания модели обозначены желтым цветом,
ground-truth разметка – зеленым

Сравнение с существующими методами

Модель	Precision	Recall	F-score
Collabonet	79.70	77.47	78.56
BioBERT	84.32	85.12	84.72
Wang et al. 2018	81.11	78.91	80.00
GMNet	**81.59**	**79.13**	**80.34**

Модель	Precision	Recall	F-score
tmVar	94.96	79.01	86.25
nala	86.32	92.20	89.16
SETH	96.42	74.66	84.15
GMNet	**87.71**	**86.48**	**87.09**

Сравнение результатов показателей качества существующих современных моделей с GMNet

Выносится на защиту

Полученные результаты

- Разработана нейросетевая модель для решения задачи извлечения именованных сущностей генов и мутаций
- Модель дает качество, сравнимое с существующими современными методами
- Проведенные вычислительные эксперименты показывают состоятельность предложенного подхода

Дальнейшие исследования

- Использовать дополнительную информацию на уровне n-gramm
- Использовать дополнительные признаки, такие как часть речи и регистр слова