Министерство науки и высшего образования Российской Федерации
Московский Физико-технический институт
(Государственный Университет)
Факультет управления и прикладной математики
Кафедра Интеллектуальных систем
при Вычислительном Центре им. А.А.Дородницына РАН

Выпускная квалификационная работа "Разработка и реализация нейросетевой системы для извлечения именованных сущностей генов и мутаций из медицинских текстов"

Студента 4-го курса Сотникова Антона Дмитриевича

Научный руководитель
д.т.н., Дулин С.К.

Научный консультант
д.т.н., Хорошевский В.Ф.

Москва, 2020

Аннотация

Решается задача распознавания именованных сущностей генов и мутаций в биомедицинских текстах. Предлагается система GMNet, основанная на рекуррентных нейронных сетях с использованием двунаправленной долгой краткосрочной памяти и условных случайных полей. Модель получает на вход последовательность токенов и на выходе определяет класс каждого токена. GMNet обучается на нескольких данных, находящихся к открытом доступе. Система, представленная в данной работе, достигает 80.34% и 87.09% по F-мере для сущностей генов и мутаций, соответственно, что является соизмеримым с результатами существующих моделей и приводит к выводу о ее высокой производительности при решении задачи извлечения именованных сущностей из медицинских текстов.

Ключевые слова: *распознавание именованных сущностей, рекуррентные нейронные сети, случайные условные поля, гены и мутации.*

Содержание

1 Введение

В настоящее время наблюдается значительный рост биомедицинских исследований и связанных с ними публикаций. Поэтому извлечение важной и ценной информации из них становится всё более сложной задачей. Исследователи проделывают серьёзную работу по поиску источников информации о биологических и медицинских сущностях, их структуризации и дальнейшей разметки [6]. Всё это делается для повышения качества новых исследований. Технически, поиск имен генов и мутаций является поиском именованных сущностей в тексте (NER), но существует ряд обстоятельств, сильно усложняющих решение такой задачи: огромное число генов и мутаций, которое растет с каждым днем; не существует единой записи подобных сущностей – они могут быть записаны в стандартном (*g.3912G>C*), полустандартном форматах (*3992-9g->a mutation*) или полностью на естественном языке (*deletion of 10 and 8 residues from the N- and C-terminals*). Также названия генов и мутаций могут встречаться в тексте и с другими биомедицинскими сущностями, которые имеют сходную морфологию и сходный контекст. Все это делает поиск рассматриваемых сущностей трудной и важной задачей. Именно поэтому так важно создавать быстрые и точные методы обработки естественного языка и анализа биомедицинской литературы, чтобы извлекать полезную информацию из возрастающего потока данных.

Цели и задачи данной работы. Основной целью исследования является построение модели на основе рекуррентных нейронных сетей и достижения после её реализации соизмеримого или более высокого качества выделения именованных сущностей генов и мутаций по сравнению с уже существующими моделями. Для реализации этой цели были поставлены следующие задачи:

- изучить существующие подходы к решению задачи извлечения именованных сущностей генов и мутаций;
- провести поиск существующих размеченных корпусов, пригодных для задач машинного обучения, для решения этой задачи;
- определить пригодность найденных источников данных;
- провести предобработку данных;
- реализовать основные алгоритмы и архитектуру нейронной сети;

- модернизировать модель путем обучения дополнительных векторных представлений для каждого символа с целью увеличения информации о токенах, не входящих в словарь (Out Of Vocabulary problem);
- реализовать модернизированную модель, обучить её на имеющихся данных, провести поиск оптимальных гиперпараметров;
- провести вычислительный эксперимент и получить значения метрик качества.

Научная новизна. Используется новый подход в решении задачи извлечения именованных сущностей генов и мутаций из медицинских текстов, основанный на использовании контекстной информации с помощью рекуррентных нейронных сетей.

Методы исследования. Использованы методы машинного обучения и нейронных сетей, методы классификации отдельных лексем (токенов) в текстах.

Практическая ценность. Полученная модель может быть использована в качестве встраиваемого модуля в более общие модули. Например, с её помощью можно решать следующие задачи.

1. Строить временные ряды по интересам авторов по разным генам и мутациям. Это дает тренды в данной области и позволяет делать предположения о том, где искать новые применения полученным знаниям.
2. Строить вопросно-ответные системы с целью диагностики симптомов заболевания, вызванного той или иной генетической мутацией.

Работа состоит из пяти разделов, заключения и списка литературы. Содержание изложено на 2 страницах. Список литературы включает 26 наименования.

Во **Введении** обосновываются цели и задачи исследования, его научная и практическая значимость.

В **Разделе 2** формулируется постановка задачи, а также используемые для оценки ее решения показатели качества.

В **Разделе 3** приводится анализ существующих методов решения задачи извлечения именованных сущностей генов и мутаций из биомедицинских текстов.

В **Разделе 4** описывается архитектура предлагаемой в работе модели.

В **Разделе 5** описываются используемые данные, параметры обучения модели, вычислительный эксперимент и анализ полученных результатов.

В **Заключении** фиксируются основные результаты работы и указываются направления дальнейших исследований.

2 Постановка задачи

Для определения меток генов и мутаций в задаче рассматривается подход **sequence labeling** (разметки последовательности), являющийся в данном случае обобщением метода классификации. Рассматриваются множество $\mathcal{K}$, состоящее из 5 классов (меток, тэгов) – O, B-GENE, I-GENE, B-MUT, I-MUT, где префиксы B и I означают начало (begin) и внутренние составляющие (inside) именованной сущности, а O – остальное (other).

2.1 Формальная постановка задачи

Дана выборка и множество меток

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

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

Требуется построить модель

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

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

Функция ошибки

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

Решается задача оптимизации:

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

2.2 Показатели оценки качества

Для оценки качества классификации токенов в задаче NER используются метрики *Precision, Recall, F1-measure* [3]. Для их вычисления вводятся следующие обозначения:

- **Correct (COR)** – совпадение метки;
- **Incorrect (INC)** – несовпадение метки;
- **Partial (PAR)** – границы прогнозируемой сущности перекрываются с ground-truth границами;

- **Missing (MIS)** – сущность не была найдена;
- **Spurius (SPU)** – сущность была найдена неверно.

Далее вводятся число всех возможных сущностей в корпусе

$$Possible(POS) = COR + INC + PAR + MIS = TP + FN$$

и число сущностей, предсказанных моделью

$$Actual(ACT) = COR + INC + PAR + SPU = TP + FP.$$

Возможны два сценария расчета метрик точности и полноты в зависимости от учета предсказанных границ именованных сущностей.

Строгое соответствие не учитывает те предсказания, которые полностью не совпадают с ground-truth границами. В таком случае метрики принимают вид

$$Precision = \frac{COR}{ACT} = \frac{TP}{TP + FP}, \qquad Recall = \frac{COR}{POS} = \frac{TP}{TP + FN}.$$

Частичное совпадение штрафует за несовпадение границ, однако учитывает, что сущность была обнаружена почти верно:

$$Precision = \frac{COR + 0.5 \cdot PAR}{ACT}, \qquad Recall = \frac{COR + 0.5 \cdot PAR}{POS}.$$

Зная значения точности и полноты для предсказаний модели, можно посчитать их среднее гармоническое:

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

В данной работе предлагается использовать сценарий строгого соответствия.

3 Обзор существующих методов

В данном разделе будут более подробно рассмотрены конкретные методы, с помощью которых решалась рассматриваемая задача.

Прежде всего нужно определить термины, которые являются наиболее ценными для проводимого исследования. Такие термины называются именованными сущностями, а задача по их нахождению (Named Entity Recognition, NER) является одной из самых важных для автоматического поиска этих объектов в тексте и дальнейшей их классификации к определённым типам. Чаще всего задача NER является первым этапом в решении более общих задач, например, при поиске отношений между сущностями или построении вопросно-ответных систем, так как она позволяет получить ценную информацию из текстов, что существенно помогает в исследованиях. В биомедицине наибольший интерес представляют именованные сущности генов, белков, их мутаций, химических соединений, а также заболеваний [9]. Тем не менее, задача NER в биомедицинских текстах является довольно сложной, так как объекты могут состоять из смеси различных символов или нескольких слов. Кроме того, имеют место сокращения и синонимы. Для ее решения существуют два подхода: на основе обучения по прецедентам и без учителя.

3.1 Методы обучения без учителя

3.1.1 Методы, основанные на словарях

Первые подходы к решению задачи NER в области биомедицины основывались на использовании правил и словарей (rule-, vocabulary-based подходы) [26]. Использование словарного метода подразумевает составление некоторого исходного словаря, на основе которого происходит извлечение биологических сущностей из текста.

Так, в [13] предлагается использовать следующий подход. Имена генов выгружаются из банка названий генов и белков, который являлся словарем. Да-

A	AAAC	**E**	AACG	**I**	AAGT	**M**	ACAC	**Q**	ACCG	**U**	ACGT	**Y**	AGAG
B	AAAG	**F**	AACT	**J**	AATC	**N**	ACAG	**R**	ACCT	**V**	ACTC	**Z**	AGAT
C	AAAT	**G**	AAGC	**K**	AATG	**O**	ACAT	**S**	ACGC	**W**	ACTG		
D	AACC	**H**	AAGG	**L**	AATT	**P**	ACCC	**T**	ACGG	**X**	ACTT		

0	AGCC	**4**	AGGG	**8**	AGTT	**/**	ATCC	**,**	ATCC	**?**	ATCC	**-**	ATCC
1	AGCG	**5**	AGGT	**9**	ATAT	****	ATCC	**;**	ATCC	**"**	ATCC		
2	AGCT	**6**	AGTC	**]**	ATCC	**(**	ATCC	**:**	ATCC	**.**	ATCC		
3	AGGC	**7**	AGTG	**[**	ATCC	**)**	ATCC	**!**	ATCC	**space**	ATCC		

Рис. 1: Таблица перевода в нуклеотидные цепочки.

лее, часть слов фильтруется за исключением часто встречающихся анлийских слов и записей, состоящих только из цифр. Затем имена преобразовываются в нуклеотидные последовательности путем замены каждого символа заранее определенной уникальной нуклеотидной комбинацией с помощью специальной таблицы без учета регистра. Используя такую же процедуру, в такой формат переводятся и медицинские тексты, после чего находятся полные и частичные соответствия между закодированными генами и текстом.

3.1.2 Методы, основанные на правилах

Модели, использующие rule-based подходы, распознают именованные сущности по заранее написанным правилам, определяющимся по текстовым шаблонам.

В [19] используется предобработанный словарь синонимов для извлечения потенциальных вхождений имен генов в текстах. Такой словарь связывает каждую биологическую сущность со всеми известными синонимами. После этого, на основе прописанных правил, эта система применятеся для обнаружения всевозможных вхождений имен генов на основе созданного словаря. Для устранения неоднозначности найденных совпадений генов в тексте обнаруживаются имена из внешних словарей. Эти внешние словари содержат сокращения, названия организмов, типы клеток и другие биологические объекты.

Система для ивзлечения мутаций описана в [15] и называется *MutationFinder*. Она использует регулярные выражения для идентификации названия мутации. Метод работает следующим образом: происходит разбиение текста на предложения, затем применяются регулярные выражения к каждому из них. Используя такой подход, становится возможным извлекать сущности мутаций, написанных на естественном языке. Прямым улучшением [15] является работа [22]. В ней рассматривается модель *SETH*, использующая следующие четыре модуля: номенклатура человеческих мутаций, регулярные выражения, модифицированный *MutationFinder*, а также литеральные упоминания. *SETH* может объединять найденные несколькими модулями мутации для разрешения неоднозначных ситуаций. Так, например, один модуль может обнаружить подстроку мутации, в то время как номенклатура находит полное упоминание о ней.

Подобные методы показали себя довольно надёжными и производительными и позволили сильно продвинуться в решении рассматриваемой задачи. Но есть существенный недостаток: они сильно зависят от созданных вручную

словарей и правил, которые, в свою очередь, зависят от внешних знаний, орфографических особенностей и не являются исчерпывающими для многих видов именованных сущностей.

3.2 Методы обучения с учителем

3.2.1 Модели, основанные на машинном обучении

Методы, основанные на обучении с учителем, используют размеченные данные для извлечения именованных сущностей. Точность и аккуратность разметки должны быть очень высокими, чтобы алгоритм смог правильно предсказывать метки именованных сущностей. Для устранения недостатков прошлых методов традиционные подходы к решению задачи NER были заменены на алгоритмы машинного обучения. В таких алгоритмах NER рассматривают как задачу маркировки последовательности, целью которой является нахождение наилучшей последовательности меток для входного предложения.

Так, в работе [16] были применены скрытые марковские модели и модели максимальной энтропии в качестве классификаторов для нахождения именованных сущностей генов и мутаций. Методы принимают на вход векторные представления слов, из которых состоит текущее предложение, затем каждому слову ставится в соответствие метка принадлежности к определенному классу. Но такие модели имеют существенный недостаток, называемый “смещением метки”, суть которого заключается в том, что модель отдает предпочтение меткам с большей энтропией, из-за чего начинает накапливаться ошибка.

В [25] было показано, что модель условных случайных полей (Conditional Random Fields, CRF) лишена такого недостатка. В этой публикации, помимо использования стандартных регулярных выражений, в качестве основного идентификатора был выбран классификатор на основе CRF. Это дало значительный прирост в полноте предсказанных сущностей.

Хотя большинство подходов, основанных на машинном обучении, привели к серьёзным улучшениям в решении задачи извлечения именованных сущностей генов и мутаций из медицинских текстов, они до сих пор остаются зависимыми от заранее описанных функций и человеческого труда.

3.2.2 Модели, основанные на глубоком обучении

Подходы, основанные на применении глубокого обучения и нейронных сетей [17], в последнее время всё чаще привлекают внимание исследователей, так

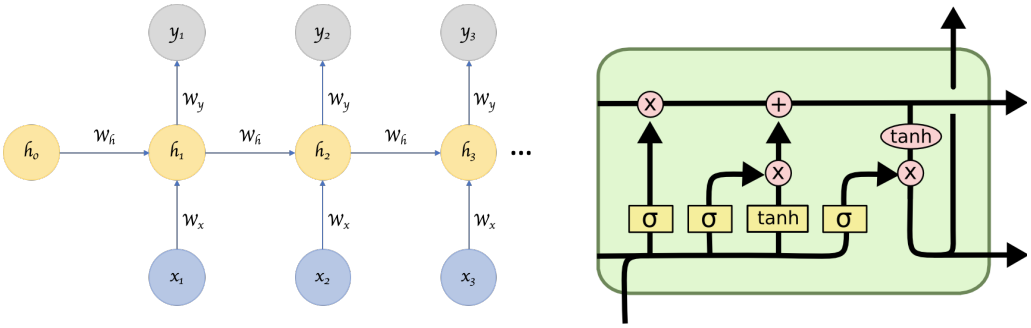
как они показывают лучшую производительность в решении таких задач обработки естественного языка, как построение языковых моделей, распознавание речи и машинный перевод. В задаче извлечения именованных сущностей также наблюдается высокий прирост в качестве при использовании нейронных сетей [17]. Так, в публикациях [11, 5, 4, 12] показана высокая эффективность моделей, использовавших долгую краткосрочную память (Long Short-Term Memory, LSTM) [10] и условные случайные поля. На вход моделей принимаются векторные представления слов (эмбеддинги), предобученные на большом объеме специализированных под предметную область текстов. Далее, эмбеддинги подаются в LSTM слой, после чего классификатор CRF присваивает очередному слову метку. Такая архитектура позволила использовать контекстную информацию и показывает наилучшие результаты по извлечению именованных сущностей и по сей день. В работе [9] предлагается использовать сверточные нейронные сети вместо слоя LSTM. Результаты оказались сравнимыми с предыдущими моделями. На сегодняшний день лучший результат демонстрирует модель BioBERT, описанная в [2]. Она основана на классической модели [1] и дообучена на большом объеме биомедицинских научных статей. BioBERT опережает по F-мере почти все существующие модели, но требует для обучения огромных вычислительных ресурсов.

Так как именно нейросетевые подходы имеют наилучшую производительность на сегодняшний день, в данной работе предлагается реализовать систему, основанную на использовании методов глубокого обучения.

4 Описание модели

4.1 Долгая краткосрочная память

Рекуррентные нейронные сети (RNN) были разработаны специально для работы с последовательной информацией (рис.1а) [17]. На каждой итерации RNN получает на вход очередное векторное представление слова $\mathbf{x_i}$. Далее на основе предыдущего внутреннего состояния $\mathbf{h_{i-1}}$ и $\mathbf{x_i}$ формируется выход $\mathbf{y_i}$ и новое внутреннее состояние $\mathbf{h_i}$. Таким образом, одна и та же матрица весов $\mathbf{W}$ последовательно применяется к очередному элементу входа, что позволяет хранить информацию о контексте каждого слова. Хотя такая модель является довольно производительной, все же она страдает от проблемы затухания и взрыва градиента [17].



а) б)

Рис. 2: (а) Пример архитектуры RNN, (б) Архитектура ячейки LSTM.

Долгая краткосрочная память (LSTM) [10] помогает в сохранении информации и решает основные проблемы классических RNN. Ячейка такой памяти состоит из трёх фильтров: фильтр забывания информации, фильтр обновления информации и выходной фильтр (рис.1б). Ниже приведены вычисления, происходящие внутри ячейки LSTM:

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-t} + \mathbf{b}_i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-t} + \mathbf{b}_f) \\
\tilde{\mathbf{c}}_t &= tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-t} + \mathbf{b}_c) \\
\mathbf{c}_t &= \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{c}}_t \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_{t-t} + \mathbf{b}_o) \\
\mathbf{h}_t &= \mathbf{o}_t * tanh(\mathbf{c}_t),
\end{aligned}
$$

где $\mathbf{c}_t$ – текущее состояние ячейки LSTM, $*$ – поэлементное умножение.

Такая структура помогает сети не забывать информацию, которая поступала достаточно давно.

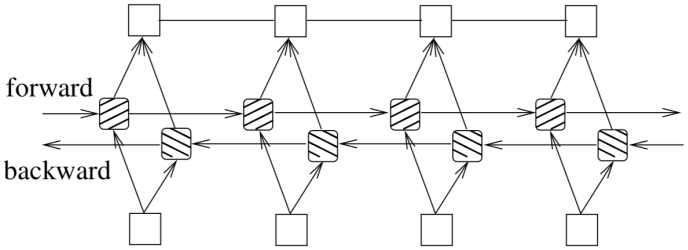


Рис. 3: Схема модели с BiLSTM.

В рассматриваемой задаче очень важно знать не только предыдущий контекст, но и будущий. Такую возможность предоставляет BiLSTM [8], основная идея которой заключается в обработке каждой последовательности в прямом и обратном направлениях в двух отдельных слоях (рис.2). Для входного предложения $(\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_n})$, содержащего $\mathbf{n}$ слов, слой LSTM вычисляет представление $\overrightarrow{\mathbf{h_i}}$ для каждого слова $\mathbf{w_i}$. Аналогично вычисляется $\overleftarrow{\mathbf{h_i}}$, для обратного предложения. В результате, BiLSTM путем конкатенации полученных двух представлений $\mathbf{h_i} = [\overrightarrow{\mathbf{h_i}}, \overleftarrow{\mathbf{h_i}}]$ получает информацию о предыдущем и будущем контекстах слова.

4.2 Условные случайные поля

В задачах разметки последовательностей, в число которых входит рассматриваемая задача, очень важно понимать зависимости между смежными метками, т.е. нужно не допускать ситуации, когда метка I-MUT возникает перед B-MUT в пределах одной сущности или возникает сразу после B-GENE. Это позволяют сделать такой класс дискриминативных статистических моделей, как условные случайные поля (рис.3) [24, 14].

На вход подается предложение $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Оценка, отражающая правильность предсказанной последовательности меток, записывается следующим образом:

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

где $\mathbf{P}_{t,y_t}$ вычисляет оценку соответствия метки y_t слову x_t (в нашем случае $\mathbf{P}_{t,y_t} = \mathbf{W}_h \cdot \mathbf{h} + \mathbf{b}_h$), $\mathbf{A}$ – матрица транзитивности (матрица вероятностей переходов от метки y_{t-1} к y_t на шаге t). Отметим, что в терминах CRF x_t называются скрытыми состояниями, а y_t – наблюдаемыми.

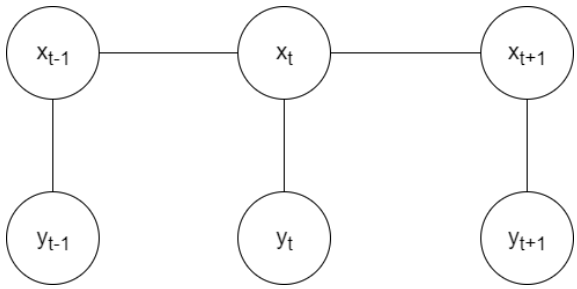


Рис. 4: Граф CRF на шаге t.

Если учесть все возможные последовательности меток, то можно получить вероятность такой последовательности:

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

$\mathbf{Y_x}$ – все возможные последовательности меток в предложении $\mathbf{x}$.

Так как в процессе обучения максимизируется логарифм правдоподобия правильной последовательности меток, то

$$\log(\mathbb{P}\{\mathbf{y}|\mathbf{x}\}) = s\left(\mathbf{x}, \mathbf{y}\right) - \log\left(\sum_{\tilde{\mathbf{y}} \in \mathbf{Y_x}} \exp(s\left(\mathbf{x}, \tilde{\mathbf{y}}\right))\right).$$

В процессе декодирования, наиболее вероятная последовательность меток y^* соответствует максимальному значению оценки s:

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

4.3 Модель GMNet

Как было сказано выше, BiLSTM способна учитывать предыдущий и будущий контекст, а CRF способны предсказывать метки сущностей, учитывая их взаимосвязь с другими метками. Используя оба этих подхода, была разработана модель GMNet для решения рассматриваемой задачи поиска именованных сущностей генов и мутаций (рис.4).

Для обучения модели используются векторные представления слов, предобученные на научных медицинских статьях, взятых из ресурса PubMed [20]. Из этих слов формируется словарь $\mathbf{V}$. Так как рассматриваемая предметная область достаточно специфична и богата редкими словами, то часть из них будет не входить в словарь, и возникнет проблема OOV (out of vocabulary) слов. Для решения такой проблемы предлагается дополнительно обучать символьные эмбеддинги, из которых состоят слова. Очередное слово разделяется на отдельные символы, которые подаются на вход BiLSTM и добавляются в словарь

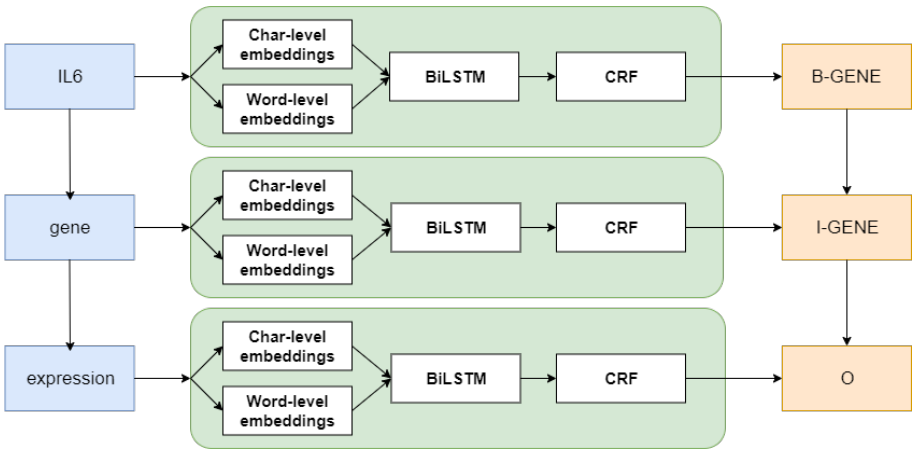


Рис. 5: Схематическая архитектура модели GMNet.

V. На выходе получаются эмбеддинги символов, которые конкатенируются к эмбеддингу слова, в которое они входят. Полученный вектор подается на вход BiLSTM, а её скрытые состояния подаются в CRF слой, который предсказывает метку текущего слова.

Модель GMNet реализована на языке Python с помощью библиотеки для работы нейронными сетями PyTorch, включает 4128462 обучаемых параметров. Размерность эмбеддингов составляет 200, используется 2 BiLSTM слоя (под символы и слова) и 1 CRF слой. Для предотвращения переобучения применяется техника дропаут с вероятностью удаления связи 0.5. В качестве оптимизатора выбран стохастический градиентный спуск (SGD) со стартовым шагом 0.001, умножающимся на 0.8 каждую эпоху.

5 Вычислительный эксперимент

В данном разделе описаны используемые в работе данные, параметры модели GMNet, а также анализ полученных результатов.

5.1 Данные

Приведем краткий перечень основных источников данных для именованных сущностей генов и мутаций.

Гены:

- *BC2GM* [7] — объемный корпус, состоящий из 20к предложений и содержащий аннотации для именованных сущностей генов и белков.
- *JNLPBA* [21] — корпус, состоящий из 22к предложений, содержащих информацию о названиях генов, белков и т.п.

Мутации:

- *tmVar* [25] — корпус, состоящий из 500 размеченных биомедицинских статей с аннотациями сущностей генных и белковых мутаций.
- *MutationFinder* [23] — корпус, включающий в себя предложения из медицинских научных статей с размеченными человеческими мутациями.

На основе выше перечисленных наборов данных был составлен корпус для обучения модели GMNet. Объем обучающей выборки – 40к, валидационной – 5.7к, тестовой – 11.4к. Обработанные данные представляют собой набор строк вида $(token, tag)$, где $token \in V$ и $tag \in$ [O, B-GENE, I-GENE, B-MUT, I-MUT].

Ниже приведена таблица 1, содержащая информацию о собранных наборах данных.

Таблица 1. Наборы данных

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

Можно заметить, что общее число сущностей мутаций меньше числа сущностей генов приблизительно в четыре раза. Соответственно, классы несбалансированы. Для решения этой проблемы используется метод, описанный в [18]. Применяется техника *oversampling*, что означает случайное добавление экземпляров «слабого» класса в выборку.

Ниже приведены примеры «сложных» и «простых» сущностей из данных.
Many erythroid cell-specific genes, including alpha and beta-globin.
Mouse interleukin-2 receptor alpha gene expression.
The -491 A to T substitution decreased the activity.
A novel heterozygous c.3703T>C change in exon.

Стоит отметить, что наибольшую трудность в поиске оказывают сущности, записанные на естественном языке (первый и второй примеры).

5.2 Эксперимент

Нейронная сеть GMNet обучалась 15 эпох. Ниже представлены результаты обучения.

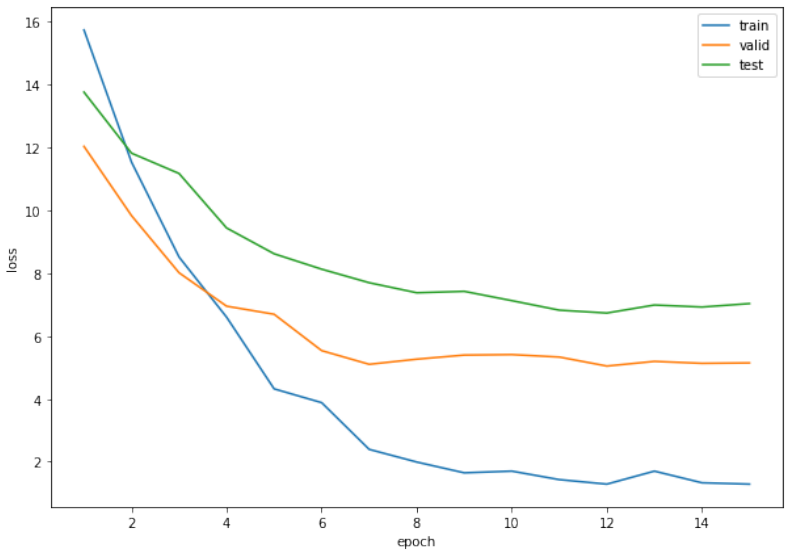


Рис. 6: График зависимости функции ошибки от числа эпох. Синий цвет соответствует обучающей выборке, оранжевый – валидационной, зелёный – тестовой.

Из рисунка 6 видно, что модель сходится, переобучения не наблюдается.

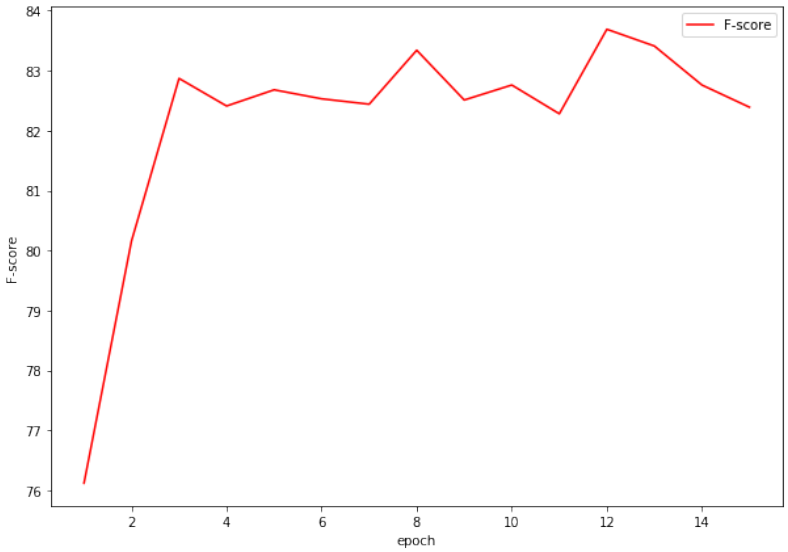


Рис. 7: График зависимости F-меры от числа эпох на тестовой выборке.

Из рисунка 7 видно, что лучший результат модель демонстрирует на 12 эпохе, далее значение показателя качества начинает снижаться и больше не достигает наилучшего результата.

В таблицах 2 и 3 содержатся значения полученных метрик *Precision, Recall, F-score* предлагаемой модели GMNet, а также существующих моделей, демонстрирующих лучшие результаты в решении рассматриваемой задачи.

Таблица 2 и 3. Результаты для сущностей генов и мутаций.

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**

Видно, что результаты получились соизмеримыми с результатами существующих решений, что говорит о состоятельности модели GMNet.

5.3 Анализ результатов

Рассмотрим полученные результаты на тестовой выборке. Примеры размеченных моделью предложений в сравнении с правильной разметкой приведены на рисунке 8.

Рис. 8: Результаты работы GMNet на тестовой выборке. Зеленым обозначена верная разметка, желтым - предсказание модели.

Как можно заметить, модель достаточно хорошо научилась находить именованные сущности генов и мутаций. Но возникают небольшие коллизии, как показано на рис.8. Видно, что модель нашла лишнюю часть в сущности гена *NF-kappaB*. Произошло это, скорее всего, из-за того, что очень часто рядом со связкой слов *DNA-Protein* возникали названия генов. Но даже такая разметка уже дает ценную информацию о гене. Также можно заметить, что редко модель выделяет сущность там, где её на самом деле нет.

6 Заключение

Была решена задача извлечения именованных сущностей генов и мутаций из биомедицинских текстов с помощью подходов глубокого обучения и использования контекстной информации отдельных сущносетй. Модель показала высокие результаты, сравнимые с уже существующими решениями. В дальнейших исследованиях предлагается:

1. использовать дополнительную информацию на уровне n-gramm;
2. использовать дополнительные признаки, такие как часть речи и регистр токена;
3. реализовать в модели механизм внимания.

Список литературы

[1] BERT: Pre-training of deep bidirectional transformers for language understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). — Minneapolis, Minnesota: Association for Computational Linguistics, 2019. — Pp. 4171–4186. https://www.aclweb.org/anthology/N19-1423.

[2] BioBERT: a pre-trained biomedical language representation model for biomedical text mining / J. Lee, W. Yoon, S. Kim et al. // *Bioinformatics*. — 2019. — sep.

[3] *Chinchor N., Sundheim B.* Muc-5 evaluation metrics // MUC. — 1993.

[4] CollaboNet: collaboration of deep neural networks for biomedical named entity recognition / W. Yoon, C. H. So, J. Lee, J. Kang // *BMC Bioinformatics*. — 2019. — may. — Vol. 20, no. S10.

[5] Deep learning with word embeddings improves biomedical named entity recognition / M. Habibi, L. Weber, M. Neves et al. // *Bioinformatics*. — 2017. — jul. — Vol. 33, no. 14. — Pp. i37–i48.

[6] *Galea D., Laponogov I., Veselkov K. A.* Exploiting and assessing multi-source data for supervised biomedical named entity recognition // *Bioinformatics*. — 2018. — Vol. 34. — Pp. 2474 – 2482.

[7] GENETAG: a tagged corpus for gene/protein named entity recognition / L. Tanabe, N. Xie, L. H. Thom et al. // *BMC Bioinformatics*. — 2005. — Vol. 6 Suppl 1. — P. S3.

[8] *Gers F., Schmidhuber J., Cummins F.* Learning to forget: Continual prediction with lstm // *Neural computation*. — 2000. — 10. — Vol. 12. — Pp. 2451–71.

[9] Gram-cnn: a deep learning approach with local context for named entity recognition in biomedical text / Q. Zhu, X. Li, A. Conesa, C. Pereira // *Bioinformatics*. — 2018. — Vol. 34. — Pp. 1547 – 1554.

[10] *Hochreiter S., Schmidhuber J.* Long short-term memory // *Neural Comput*. — 1997. — Vol. 9, no. 8. — P. 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[11] *Hong S. K., Lee J. G.* DTranNER: biomedical named entity recognition with deep learning-based label-label transition model // *BMC Bioinformatics.* — 2020. — Feb. — Vol. 21, no. 1. — P. 53.

[12] HUNER: improving biomedical NER with pretraining / L. Weber, J. Münchmeyer, T. Rocktäschel et al. // *Bioinformatics.* — 2019. — jun. — Vol. 36, no. 1. — Pp. 295–302.

[13] *Krauthammer M., Rzhetsky A., Morozov P.* Using blast for identifying gene and protein names in journal articles // *Gene.* — 2001. — 01. — Vol. 259. — Pp. 245–252.

[14] *Lafferty J.* Conditional random fields: Probabilistic models for segmenting and labeling sequence data. — 2001. — Pp. 282–289.

[15] MutationFinder: a high-performance system for extracting point mutation mentions from text / J. G. Caporaso, W. A. Baumgartner, D. A. Randolph et al. // *Bioinformatics.* — 2007. — may. — Vol. 23, no. 14. — Pp. 1862–1865.

[16] nala: text mining natural language mutation mentions / J. M. Cejuela, A. Bojchevski, C. Uhlig et al. // *Bioinformatics.* — 2017. — Jun. — Vol. 33, no. 12. — Pp. 1852–1858.

[17] Neural architectures for named entity recognition / G. Lample, M. Ballesteros, S. Subramanian et al. // *CoRR.* — 2016. — Vol. abs/1603.01360. `http://arxiv.org/abs/1603.01360`.

[18] *Padurariu C., Breaban M. E.* Dealing with data imbalance in text classification // *Procedia Computer Science.* — 2019. — Vol. 159. — Pp. 736–745.

[19] ProMiner: rule-based protein and gene entity recognition / D. Hanisch, K. Fundel, H.-T. Mevissen et al. // *BMC Bioinformatics.* — 2005. — Vol. 6, no. Suppl 1. — P. S14.

[20] Pubmed // https://pubmed.ncbi.nlm.nih.gov/.

[21] Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task / M. Huang, P. Lai, R. T. Tsai, W. Hsu // *CoRR.* — 2019. — Vol. abs/1901.10219. `http://arxiv.org/abs/1901.10219`.

[22] SETH detects and normalizes genetic variants in text / P. Thomas, T. Rocktäschel, J. Hakenberg et al. // *Bioinformatics*. — 2016. — jun. — Vol. 32, no. 18. — Pp. 2883–2885.

[23] SETH detects and normalizes genetic variants in text / P. Thomas, T. Rockt?schel, J. Hakenberg et al. // *Bioinformatics*. — 2016. — 09. — Vol. 32, no. 18. — Pp. 2883–2885.

[24] *Sutton C., Mccallum A.* An introduction to conditional random fields. — 2012.

[25] tmVar: a text mining approach for extracting sequence variants in biomedical literature / C. H. Wei, B. R. Harris, H. Y. Kao, Z. Lu // *Bioinformatics*. — 2013. — Jun. — Vol. 29, no. 11. — Pp. 1433–1439.

[26] *Ulf Leser J. H.* What makes a gene name? named entity recognition in the biomedical literature // *Briefings in Bioinformatics*. — 2005. — Vol. 6. — Pp. 357–369.