МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра дискретной математики и алгоритмики

Скидан Ольга Николаевна

АНАЛИЗ БИОИНФОРМАТИЧЕСКИХ ТЕКСТОВ МЕТОДАМИ ОБРАБОТКИ ЕСТВЕСТВЕННОГО ЯЗЫКА ДЛЯ ПРЕДСКАЗАНИЯ БЕЛКОВЫХ ВЗАИМОДЕЙСТВИЙ

Магистерская диссертация
специальность 1-31 81 09 «Алгоритмы и системы
обработки больших объемов информации»

Научный руководитель:
Тузиков Александр Васильевич,
профессор кафедры биомедицинской
информатики, член-корр.
НАН Беларусь, генеральный
директор ОИПИ НАН Беларуси,
доктор физико-математических наук

Допущена к защите
«___» _____ 2020 г.
Зав. кафедрой дискретной математики и
алгоритмики
доктор физико-математических наук,
В.М. Котов

Минск, 2020

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Извлечение белок-белковых взаимодействий является важной задачей, поскольку часто позволяют определить функции белка и те биологические процессы, в которые он вовлечен. Экспериментальные методики не позволяют с достаточной точностью предсказать факт взаимодействия между всеми, рассматриваемыми белками, поэтому требуются методы на основе компьютерного анализа данных, которые могут найти, предсказать или извлечь эти взаимодействия, исходя из свойств белков.

В настоящее время количество статей по биоинформатике, которые содержат информацию о белок-белковых взаимодействиях достаточно велико, что делает невозможным ручной просмотр всех статей. При этом не существует единого ресурса, где хранится вся информация о белковых взаимодействиях и структурах взаимодействующих белков. Существующие базы данных неполны и отличаются друг от друга, что затрудняет работу с ними. Поэтому разработка алгоритма, который из текстов научных статей может найти для двух белков информацию о наличии и особенностях этого взаимодействия, является актуальной задачей.

Глава 1
ОПИСАНИЕ ЗАДАЧИ И ПОДХОДЫ К ПОСТРОЕНИЮ СИСТЕМЫ АНАЛИЗА БИОЛОГИЧЕСКОГО ТЕКСТА

1.1 Описание обобщенной задачи

Белок-белковые взаимодействия (PPI - Protein-Protein Interaction) играют ключевую роль в различных биологических процессах. Достоверная характеристика молекулярных механизмов этих процессов требует 3D-структуры белок-белковых комплексов. Из-за ограничения экспериментальных методов, большинство структур должны быть смоделированы на основе каких-то свойств или на основе шаблона докинга. Обе парадигмы докинга производят большой объем предполагаемых моделей, и выбор правильной является нетривиальной задачей, выполняются процедуры оценивания. Часто знание нескольких обязательных остатков достаточно для успешной стыковки.

В последние годы число биомедицинских публикаций, в том числе PPI релевантных областях, быстро растет. На сегодняшний день более 16 миллионов цитат таких статей можно найти в базе данных MEDLINE[1]. Параллельно с этими источниками информации обычного текста, многие базы данных, такие как DIP[2], BIND[3], IntAct[4] и STRING[5], были созданы для хранения различных типов информации о белково-белковых взаимодействиях. Тем не менее, данные в этих базах данных в основном проверяются вручную для обеспечения их правильности и, следовательно, это ограничивает скорость передачи текстовой информацию в структуру данных, пригодную для поиска и навигации.

Получение и извлечение такой информации из литературы очень сложны из-за отсутствия формальной структуры на естественном языке в этих документах. Таким образом, автоматическое извлечение информации из биомедицинских текстов может легко обработать большие объемы биологических данных в доступных для компьютера формах. Для достижения этой цели были разработаны многие системы, такие как:

- EDGAR [6],
- BioRAT [7],
- GeneWays [8] и т. д.,

В общем случае для автоматического извлечения белково-белковых взаимодействий система должна состоять из нескольких основных модулей.

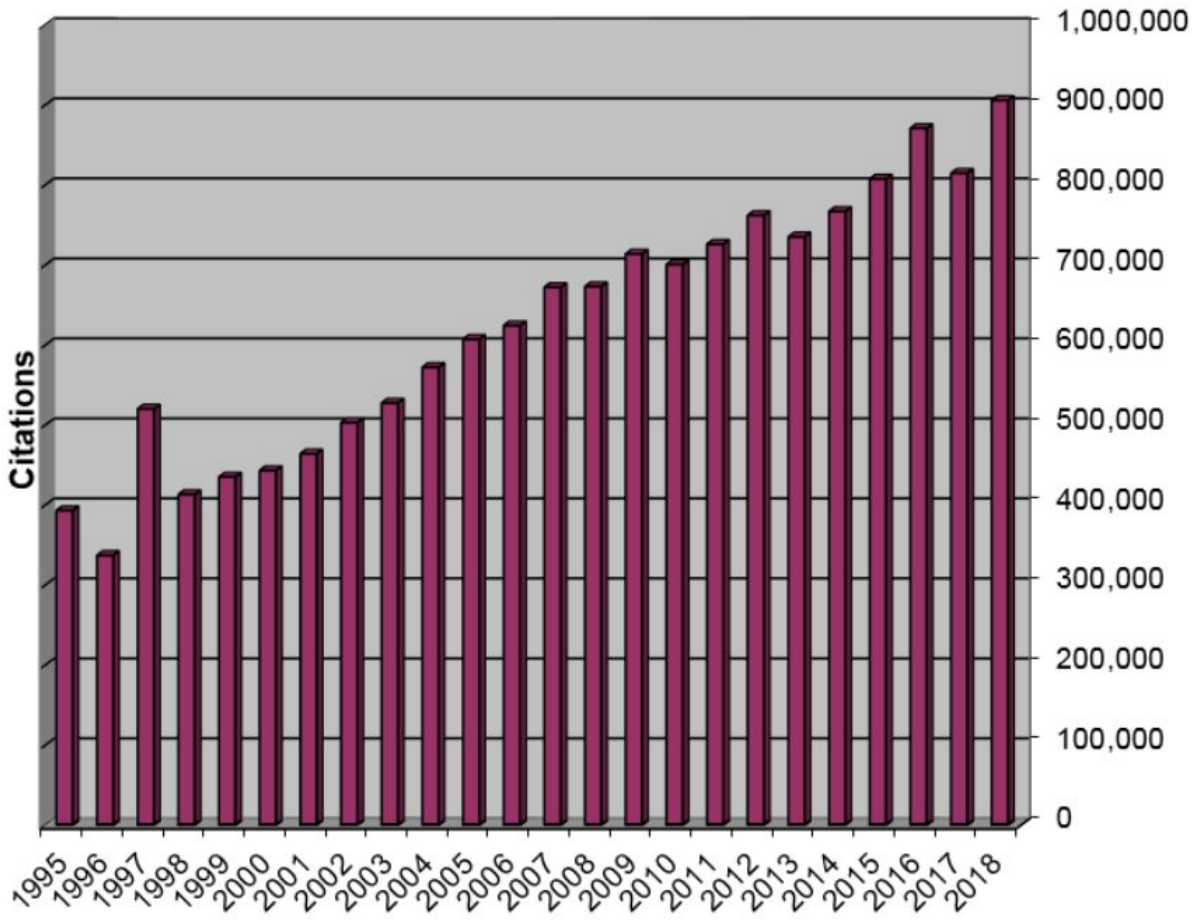

Рис.1 Количество проиндексированных ссылок, добавленных в MEDLINE с 1995г.

1.2 Модули для построения автоматического извлечения белок-белковых взаимодействий

1. Модуль ранжирования. Анализ всевозможных статей из базы данных MEDLINE представляет из себя долгую и дорогую операцию. Для ее решения используют модуль ранжирования, т.е. упорядочивание статей в порядке уменьшения вероятности, что там имеется информация о белок-белковом взаимодействии.

2. Модуль зонирования. Данный модуль разбивает документы на основные строительные блоки для последующего анализа. Типичными строительными блоками являются фразы, предложения и абзацы. В особых случаях могут быть выбраны строительные блоки более высокого уровня, такие как секции или главы. Были сравнены результаты использования различных текстовых единиц, таких как фразы, предложения и тезисов от MEDLINE, для получения информации о взаимодействии между биохимическими объектами[9]. Экспериментальные результаты показали, что тезисы, предложения и фразы могут дать сравнительные похожие результаты для извлечения релевантной информации. Однако в отношении

эффективности предложения значительно лучше, чем фразы, и примерно такие же, как тезисы.

3. Модуль распознавания имен белков. Перед экстракцией белково-белковых взаимодействий крайне важно облегчить идентификацию названий белков, которые все еще остаются сложной проблемой. Хотя сообщалось об экспериментальных результатах высоких значений точности и полноты, при пометки имен белка для конъюнктивного естественного названия встречаются несколько препятствий для дальнейшего развития [10]. Существует количественный обзор причин неоднозначности генных названий и было предложено, что исследователи могут сделать, чтобы свести к минимуму эту проблему [11].

4. Модуль экстракции белок-белкового взаимодействия. Поскольку извлечение белок-белковых взаимодействий привлекло много внимания в области извлечения биомедицинской информации, было предложено множество подходов. Решения варьируются от простых статистических методов, основанных на совпадениях генов или белков, с методами, использующими глубокий синтаксический или семантический анализ

5. Модуль визуализации. Этот модуль не так важен, как вышеупомянутые четыре модуля, но он обеспечивает дружественный интерфейс для пользователей, чтобы понять полученные результаты. Кроме того, он позволяет взаимодействовать с системой, чтобы упростить обновление базы знаний системы и в конечном итоге улучшить ее производительность.

1.3 Описание поставленных задач

Первая поставленная задача фокусируется на первом модуле для извлечения белок-белковых взаимодействий, а именно модуле ранжирования, или другими словами модуле классификации статей по наличию информации о взаимодействии.

Извлечение информации о белок-белковых взаимодействиях из литературы имеет важной значение для биомедицины, так как понимание болезней, фармакологических и других процессов требует анализа белковых сетей, образованных этими отношениями. В нескольких базах данных хранятся данные о взаимодействии белков, которые вручную обрабатываются, но, поскольку основным источником идентификации взаимодействий является научная литература, поддержание актуальности этих баз данных является сложной и дорогостоящей задачей. Было показано, что использование первого модуля, как первый шаг для распознавания именованных объектов (NER) и извлечения отношений, значительно ускоряет весь процесс извлечения нужной информации. Важным шагом в таких процессах является расстановка приоритетов или сортировка

документов для выбора статей, которые с большей вероятностью содержат соответствующую информацию.

Вторая поставленная задача относится к третьему модулю, или другими словами модулю для распознавания белков в тексте. Задача является трудоемкой из-за отсутствия единой и сложной номенклатуры названий.

С помощью двух таких модулей тексты можно классифицировать по наличию информации о взаимодействие между белками и извлекать названия белков. Конечно, такой подход не дает 100% гарантии, что если было извлечено 3 белка, то они друг с другом взаимодействуют, но может помочь для нахождения информации о новых белках или для последующей ручной обработке.

Глава 2
ИСПОЛЬЗУЕМЫЕ ДАННЫЕ

2.1 Используемые данные

Были использованы данные из соревнования BioCreative III для классификации текстов. Данный корпус состоит из вручную аннотированных MEDLINE статей, содержащий 2280 документов для тренировочных данных, 4000 - для валидационных и 6000 - для тестовых. Тренировочные данные содержат одинаковое число положительных и отрицательных примеров, в то время как валидационный и тестовые наборы очень не сбалансированы (около 15-17% положительных примеров, что отражает реальную картину).

	Кол-во статей	Количество статей, сод. информацию о ББВ	Количество статей, несод. информацию	% статей, сод. информацию о ББВ	Количество различных журналов
Тренировочные данные	2,280	1,140	1,140	50%	118
Валидационные данные	4,000	682	3,318	17.05%	113
Тестовые данные	6,000	910	5,090	15%	112
Всего	12,280	2,732	9,548		121

Таб.1 Используемые данные

Для задачи распознавания имен белков были использованы два набора данных: JNLPBA и BC2GM, которые были построен на базе MEDLINE тезисов. Теги для данных наборов были представлены в формате записи BIOES для именованных имен сущностей [16]. Данные наборы фокусируются на объектах ген / белок. Информация о генах была отфильтрована.

В то время как набор данных JNLPBA имеет только данные для обучения и тестирования, BC2GM содержит даные для обучения, валидации и тестирования. Для JNLPBA использовалось часть его обучающего набора в качестве его валидационных данных, размер которого совпадает с размером его тестового набора равный 20%.

Название датасета	Количество предложений	Количество аннотированных белков
JNLPBA	22 562	35 336
BC2GM	20 510	24 583

Таб. 2. Количество предложений для нейросети и количество аннотированных белков в них.

2.2 Структура MEDLINE и индексация статей

Как базу биомедицинских статей будем использовать MEDLINE. MEDLINE является ведущей библиографической базой данных Национальной медицинской библиотеки США, которая содержит более 25 миллионов ссылок на журнальные статьи в области наук о жизни с уделением особого внимания биомедицине. Отличительной особенностью MEDLINE является то, что записи индексируются с помощью медицинских предметных рубрик.

Каждая запись MEDLINE представляет собой одну статью из научной литературы. Идентификатор Medline представляет собой целое число, которое однозначно идентифицирует запись. Если в содержание записи вносятся исправления, идентификатор пользователя не изменяется. MEDLINE uid - это самый простой и надежный способ идентифицировать запись. Месяц записи - это месяц и год, когда запись стала частью публичного представления Medline. Это не то же самое, что дата публикации статьи. Это в основном полезно для отслеживания того, что опубликовано нового со времени предыдущего запроса Medline.

2.3 Векторное представления слов

Для текстовых задач входные данные должны быть закодированы так, чтобы их можно использовать, как вход глубокой нейронной сети. Это может быть достигнуто путем представления каждого слова в качестве вектора относительно небольшого размера. Таким образом, каждый документ представлен последовательностью векторов слов, которые подаются непосредственно в сеть.

Далее будет использован метод векторного представления Word Embeddings. Word Embeddings - это метод для векторного представления слов на основе больших аннотированных корпусов. Слова с похожей семантикой данный метод представляет векторами близкими в векторном пространстве. Использование таких представлений вместе с методами глубокого обучения

приводит к улучшению результатов в различных задачах NLP, включая устранение неоднозначности слов, классификацию текстов и распознавание именованных сущностей .

В качестве Word Embeddings будем использовать вложения, заранее натренированные на PubMed (PM), PubMed Central(PMC) и Википедии с использованием алгоритма Word2Vec и фреймворка gensim. Данные векторные представления были натренированы на следующем корпусе - 15 млн. абстрактов полной базы данных MEDLINE, содержащих 775 различных слов. Размер векторного представления - 200 компонент.

Глава 3
МОДЕЛЬ КЛАССИФИКАЦИИ ТЕКСТОВ

3.1 Нейросетевые архитектуры

Сверточные нейронные сети (CNN) являются одной из самых популярных сетевых архитектур, используемых в глубоком обучении, которые широко применяются в задачах распознавания и классификации изображений с очень хорошей производительностью. Различные работы также демонстрируют их применение в задачах классификации текста. Тем не менее, последовательная природа природных текстов может быть лучше смоделирована рекуррентными нейронными сетями (RNN), которые содержат петлю обратной связи, которая позволяет сети использовать информацию, касающуюся предыдущего состояния.

Сети с кратковременной памятью (LSTM) представляют собой особый тип RNN, в котором вводится набор информационных шлюзов, которые позволяют этим сетям изучать долгосрочные зависимости, избегая проблемы затухающего градиента.

Важное соображение при определении глубокой нейронной сети для проблемы классификации связано с выбором топологии сети, а именно типа и количества слоев, количества единиц в каждом слое, параметров модели, таких как функция активации для каждого слоя, функции потерь и алгоритм оптимизации.

Еще один важный аспект связан с переобучением, что означает, что сеть способна выучить «лучшее» представление для данных, используемых при обучении, но не может обобщаться для невидимых данных. Существуют различные стратегии и обычно используются для решения этой проблемы следующее, а именно ранняя остановка, отсев и регуляризация. Первый основан на остановке тренировки, когда значение функции потери, измеренное на валидационной части данных, перестает уменьшаться. Отсев замораживает веса некоторых ребер в соответствии с прошлой итерацией. Это означает, что после итерации обучения веса этих единиц не изменяются, так что сеть вынуждена улучшать результат изменяя другие веса и не фокусируется на некоторых частях пространства функций. Регуляризация - это общая стратегия, используемая для того, чтобы избежать переобучения.

3.2 Решение поставленной задачи

Первым уровнем предложенного решения являются эмбеддинги слов, заранее натренированные на PubMed (PM), PubMed Central(PMC) и Википедии с использованием алгоритма Word2Vec и фреймворка gensim. Слой эмбеддингов сопровождается слоем отсева, чтобы уменьшить

чрезмерную подгонку и тем самым улучшить обобщение. Конечная сеть использует коэффициенты отсева 0,2 и 0.3 на разных уровнях, коэффициент 0.2 используется после эмбеддингово слоя, 0.3 - во всех остальных. Слой отсева используется после слоя, отличного от самого отсева.

Далее следует слой LSTM, который был настроен на 100 единиц. Слой отсева также использовался после слоя с ячейкой памяти со значением 0,3. Выход LSTM затем обрабатывается слоем пулинга и подключается к полносвязному слою, который состоит из 20 нейронов с функцией активации ReLU. Затем идет финальный слой, который состоит из одного нейрона, на которые применяется функция активации - сигмоида - для получения вероятностных выходов в диапазоне 0–1 для каждого текста.

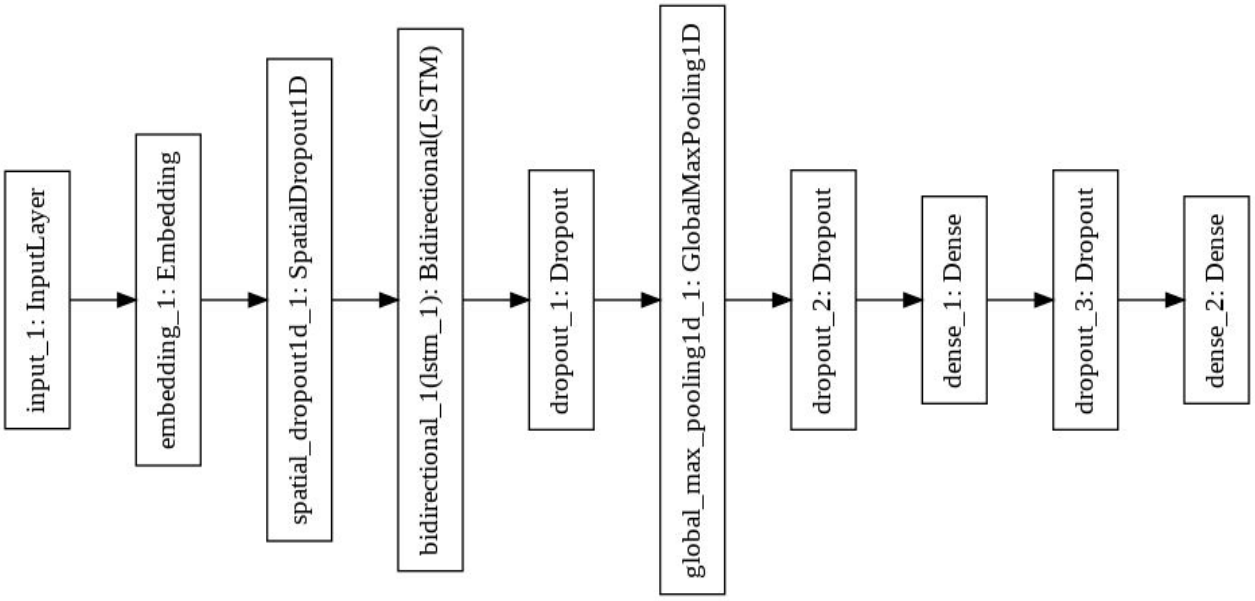

Рис.2 Нейросетевая архитектура для решения поставленной задачи

Полученная архитектура проиллюстрирована на Рис. 2. Кроме этого данная архитектура была реализована с ячейкой памяти GRU, но результаты для данной сети оказались хуже, чем для архитектуры с Рис. 2. Также результаты оказались хуже для всех других реализованных архитектур с различными конфигурациями и параметрами.

Глава 4
ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ И СРАВНЕНИЕ С СУЩЕСТВУЮЩИМИ КЛАССИФИКАТОРАМИ

4.1 Численная оценка качества классификатора

В простейшем случае такой метрикой может быть доля объектов по которым классификатор принял правильное решение.

$$Accuracy = \frac{P}{N} \qquad (1)$$

где, P – количество объектов по которым классификатор принял правильное решение, а N – размер обучающей выборки.

Точность (precision) и полнота (recall) являются метриками которые используются при оценке большей части алгоритмов классификации. Иногда они используются сами по себе, иногда в качестве базиса для производных метрик, таких как F-мера или R-Precision.

Точность системы в пределах класса – это доля объектов действительно принадлежащих данному классу относительно всех объектов которые система отнесла к этому классу. Полнота системы – это доля найденных классификатором объектов принадлежащих классу относительно всех объектов этого класса в тестовой выборке.

Эти значения легко рассчитать на основании таблицы контингентности, которая составляется для каждого класса отдельно.

Категория		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

Таб.2 Таблица контингентности

В таблице 2 содержится информация сколько раз система приняла верное и сколько раз неверное решение по документам заданного класса. А именно:

TP — истинно-положительное решение;

TN — истинно-отрицательное решение;

FP — ложно-положительное решение;

FN — ложно-отрицательное решение.

Тогда, точность и полнота определяются следующим образом:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю :

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

4.2 Полученные результаты и сравнительный анализ

Т.к. использованные данные были использованы из соревнования BioCreative III, сравним лучшие результаты соревнования с результатами, полученными с помощью архитектуры Рис. 2.

	Accuracy	F1-score	MCC
Своя модель	0.863	0.609	0.513
Топ-1	0.887	0.614	0.550
Топ-2	0.876	0.603	0.530
Топ-3	0.868	0.323	0.334
Топ-4	0.824	0.561	0.481

Таб.3 Полученные результаты и их сравнение с результатами соревнования

Кроме этого на Рис. 3, Рис. 4 и Рис. 5 приведены кривые обучения и ROC-кривая, показывающая, как хорошо мы обучились.

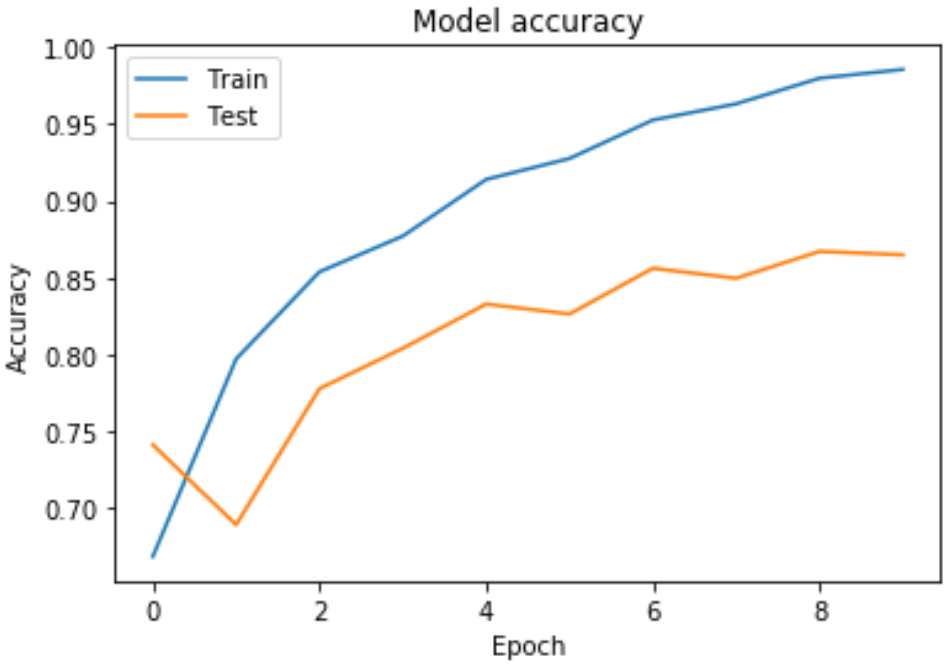

Рис.3 График точности на тренировочном и валидационном наборе данных для каждой эпохи

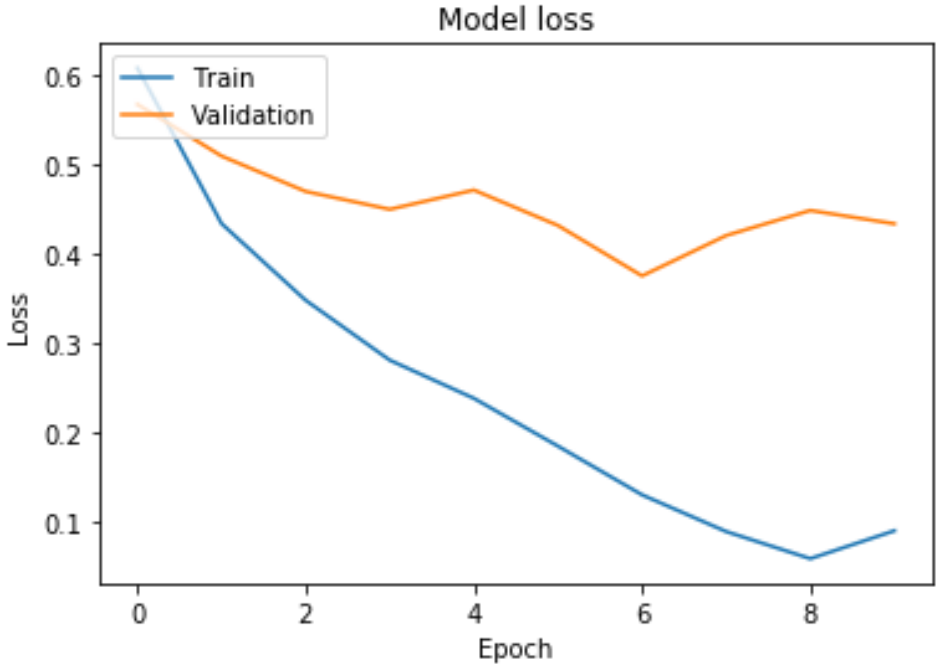

Рис.4 График ошибки на тренировочном и валидационном наборе данных для каждой эпохи

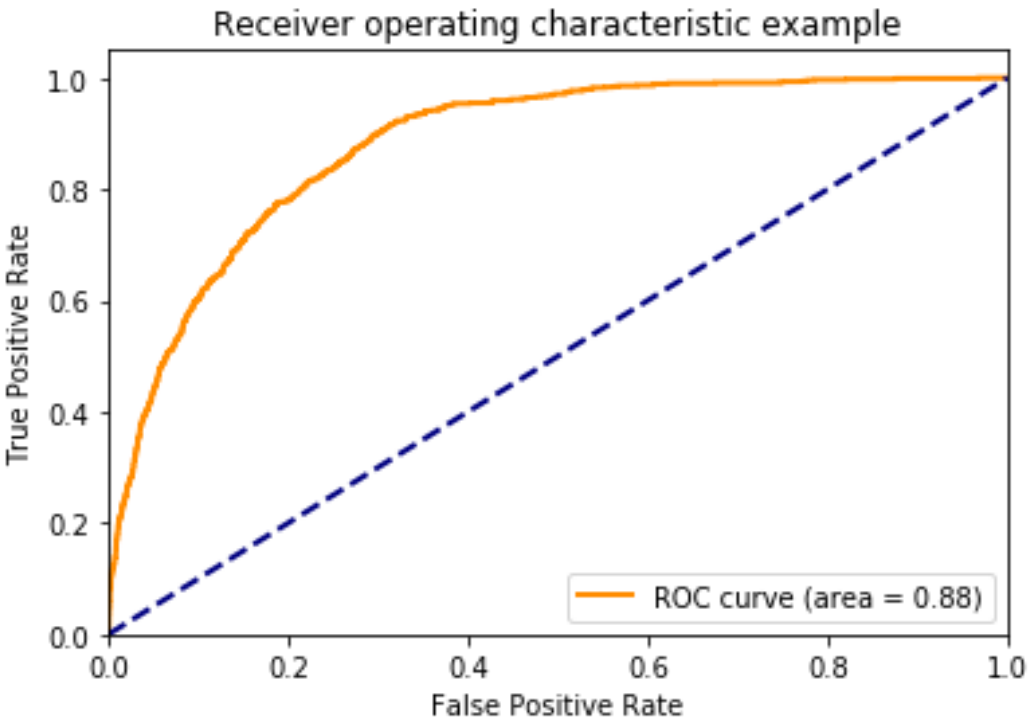

Рис.5 ROC-кривая для поставленной задачи бинарной классификации

Модель часто ошибается в двух случаях: на статьях, где есть упоминание белков, но нет информации о взаимодействие; на статьях, где явно есть слова о каком-либо взаимодействие. Для получения лучших результатов необходимо иметь больше размеченных данных, как видно из Рис. 4 уже на четвертой эпохе модель начинается переобучаться и происходит ранняя остановка модели.

Глава 5
НАХОЖДЕНИЕ И ИЗВЛЕЧЕНИЕ БЕЛКОВ ИЗ БИОЛОГИЧЕСКИХ ТЕКСТОВ

5.1 Особенности названий белков

Названия белков могут быть классифицированы по следующим трем категориям из их структуры.

1. Отдельные слова с заглавными буквами, цифрами и не алфавитными буквами. В основном происходит от имени гена (напр. Nef, p53, Vav)

2. Составные слова с заглавными буквами, цифрами и не алфавитными буквами. (напр. interleukin 1 (IL-1) - чувствительная киназа)

3. Одно слово, состоящее только из строчных букв (например, актин, тубулин, инсулин)

В области медицины и биологии часто сообщается о новых находках белков и функций. Когда новый белок обнаружен, обычно создается новый термин, который можно четко отличить его от других белков. Следовательно, в реальных работах названия белков типа 3 встречаются относительно редко, тогда как названия белков, такие как типы 1 и 2, часто встречаются. Эти названия, как правило, являются новыми словами, которые вводятся исследователем/группой и которые обновляются ежедневно.

В дополнение к недавно введенным словам, вариация и непоследовательность в ссылках на уже известные материалы является еще одной серьезной проблемой. Особенно при упоминании белка, название которого объясняет его роль, выражение может быть почти произвольным. Кроме того, собственные слова могут иметь несоответствие в написании.

1. Авторы часто используют оригинальные слова вместо сокращений, меняют регистры букв и игнорируют неявные правила генерации имен.

- рецептор эпидермального фактора роста(epidermal growth factor receptor) или рецептор EGF или EGFR
- циклиновый комплекс D1-cdk4 или циклиновый комплекс D1-Cdk4
- c-jun, c jun или c-Jun.

2. Ниже название объясняет белковую функцию.

- Ras guanine nucleotide обменный фактор Sos

- Ras guanine nucleotide высвобождающий белок Sos
- Ras обменник Sos
- GDP-GTP обменный фактор Sos
- Sos (mSos), белок обмена GDP/GTP для Ras

Они показывают, что описание названий белков в области медицины и биологии чрезвычайно разнообразны.

3. Следующие примеры включают предлог и/или соединение. Из-за неоднозначности зависимостей изменение в описании может быть более сложным по сравнению со вторыми примерами.

- альфа-субъединица p85 PI 3-киназы
- SH2 и SH3 домены Src

Таким образом, описание названий белков часто зависит от стиля автора, и нет никакой гарантии, что один и тот же белок появится в одном и том же описании в разных предложениях и документах.

Несмотря на эту произвольность описания названия белка, в технических терминах этой области существует особые характеристики. Характерные слова, содержащие заглавные буквы, цифры и специальные символы, как подчеркнуто в следующих примерах, часто наблюдаются в биоинформатических текстах. Эти слова можно четко отличить от общих слов.

- Домены гомологии ***Src*** (***SH***) 2 и ***SH3***
- ***p54 SAP*** киназа

Эти слова предоставляют читателю большое количество информации и могут рассматриваться как основа названий белков. В связи с этим можем называть такие слова, которые появляются в названиях белков, “ключевыми словами”.

Кроме того, как далее показано, могут быть включены ключевые слова, которые описывают функцию и символы составного слова.

- ***Рецептор*** EGF
- Ras GTPase-активирующий ***белок*** (GAP)

Мы будем называть такие слова “f-терминами” (характерными терминами).

Сосредоточив внимание на этих характеристиках, станет легче находить кандидатов названий материалов, в том числе тех, которые были недавно представлены.

5.2 Методы нахождения названия белков

Распознавание именованных объектов (NER - name entity recognition) - это компьютеризированная процедура распознавания и маркировки объектов в заданных текстах. В биомедицинской области типичные типы объектов включают заболевания, химические вещества, гены и белки. Распознавание биомедицинских названных объектов (BioNER) является важным строительным блоком многих последующих приложений для анализа текста, таких как извлечение взаимодействий между белками.

NER в области биомедицинского интеллектуального анализа текста в основном сосредоточен на подходах, основанных на использовании словарей, правил и машинного обучения. Словарные системы имеют простую и интуитивно понятную структуру, но они не могут обрабатывать объекты вне словаря или многозначные слова, что приводит к низкому значению полноты. Кроме того, создание и поддержание всестороннего и современного словаря требует значительного объема ручной работы.

Подход, основанный на правилах, является более масштабируемым, но для его работы к набору данных необходимы специально созданные наборы функций, которые также задаются вручную. Основанные на правилах и словарях подходы могут достигать высокой точности, но также могут давать неправильные предсказания, когда в предложении появляется новое слово, которого нет в обучающих данных (проблема вне словарного запаса). Проблема вне словарного запаса часто возникает, особенно в биомедицинской области, поскольку в этой области зачастую регистрируются новые биомедицинские термины, такие как названия новых генных структур(белков).

В последнее время исследования продемонстрировали эффективность методов глубокого обучения, в том числе эффективность рекуррентных нейронных сетей (RNN) для NER в биомедицинском тексте[12], состоящих из двунаправленных сетей краткосрочной памяти (BiLSTM) и условного случайного поля (CRF). Кроме этого было показано, что при использовании векторного представления символов для охвата характеристик, таких как орфографические особенности биомедицинских объектов, достигли современного уровня производительности, демонстрируя эффективность векторного представления слов на уровне символов в BioNER[13].

5.3 Получение векторного представления символов

Чтобы предоставить модели морфологическую информацию на уровне символа (например, «-ase» является обычной для белковых сущностей), следует использовать информацию на уровне символа для каждого слова. Для построения вложения слов на уровне символов (CLWE), используем сверточную нейронную сеть(CNN).

Слово w_t, состоящее из M символов M, представим в виде $w_t = \left\{ c_1^t,\ c_1^t,\ ...,\ c_M^t \right\}$, где $c_i^t \in R^{d^{char}}$ - это случайно инициализированное вложение символа для каждого уникального символа. Для CNN в каждое слово должно быть добавлено до и после ((k - 1) / 2) пустых символов в соответствии с размером окна k. Получаем вектор для окна C_i^t, просто сливанием вложений символов c_i^t с вложениями символов (k - 1) / 2 символов с обеих сторон:

$$C_i^t = \left[c_{i-(k-1)/2}^t, ..., c_i^t, ...,\ c_{i+(k-1)/2}^t \right] \in R^{k*d^{char}} \quad (5)$$

Для вектора окна C_i^t мы выполняем операцию свертки следующим образом:

$$\left[x_t^c \right]_j = max_{1 \le i \le M} \left[W_{char} * C_i^t + b_{char} \right]_j \quad (6)$$

где $W_{char} \in R^{d^{clwe} * kd^{char}}$ и $b_{char} \in R^{d^{clwe}}$ обозначают обучаемый фильтр и смещение соответственно.

Получаем поэлементные максимальные значения, и на выходе получается вложение слова уровня символов, обозначенное как $x_t^c \in R^{d^{clwe}}$. Для дальнейшего обучения модели объединим вложение слов на уровне символов с обычным векторным представление слов, обученных на биомедицинских корпусах, как $\widehat{x}_t^c = \left[x_t, x_t^c \right]$.

В дальнейшем данные вложения будут использоваться, как часть общей рекуррентной сети для извлечения белков в рамках научно-исследовательской работы. Также для тренировки символьных вложений была реализована и протестирована архитектура с RNN и ячейкой памяти LSTM.

Конечные результаты моделей, полученные с помощью различных методов тренировки символьных вложением, различаются незначительно поэтому являются взаимозаменяймыми.

5.4 Решение поставленной задачи

Первым уровнем входных данных предложенного решения являются эмбеддинги слов, заранее натренированные на PubMed (PM), PubMed Central(PMC) и Википедии с использованием алгоритма Word2Vec и фреймворка gensim. Далее за слоем эмбеддингов слов следует TimeDistributed слой, для организации последовательности слов.

Второй уровень входных данных - эмбеддинги символов, которые были описаны в предыдущем разделе на основе сверточных нейронных сетей и фильтров различной длины. После этого два входных уровня конкатенируются и сопровождаются слоем отсева, чтобы уменьшить чрезмерную подгонку и тем самым улучшить обобщение. Конечная сеть использует коэффициент отсева 0,3.

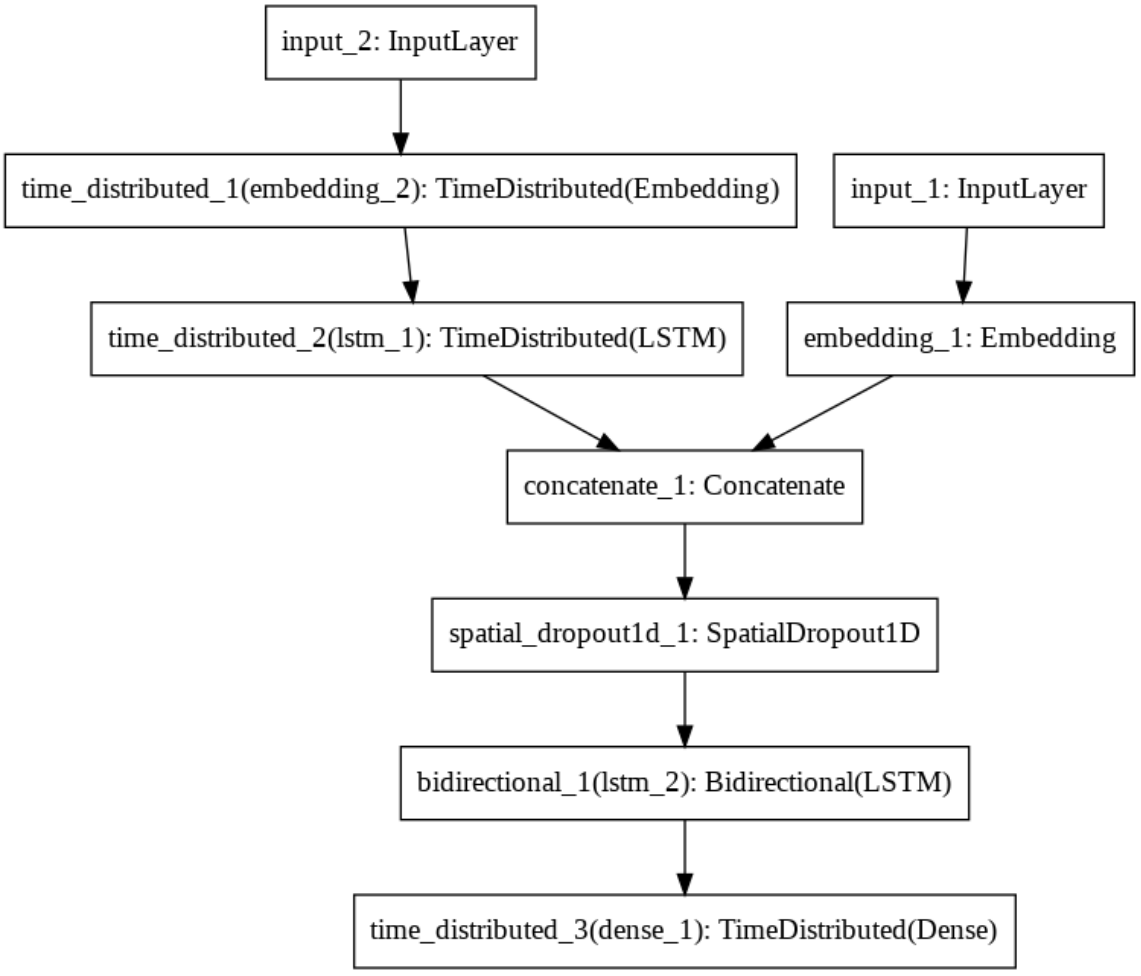

Рис.6 Нейросетевая архитектура для решения поставленной задачи

Далее следует слой двунаправленного LSTM, который был настроен на 30 единиц. Выход LSTM затем обрабатывается слоем TimeDistributed, на

которые применяется функция активации - сигмоида - для получения вероятностных выходов в диапазоне 0–1 для каждого слова.

Полученная архитектура проиллюстрирована на Рис. 6. Кроме этого данная архитектура была реализована с ячейкой памяти GRU и однонаправленным LSTM , но результаты для данной сети оказались хуже, чем для архитектуры с Рис. 6.

Глава 6
ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ И СРАВНЕНИЕ С СУЩЕСТВУЮЩИМИ КЛАССИФИКАТОРАМИ

6.1 Полученные результаты и сравнительный анализ

На таких же данных был продиктован алгоритм и описан в статье - Wang et al.(2018) STM [17]. Сравним лучшие результаты полученные в статье с результатами, полученными с помощью архитектуры Рис. 4 на основе тестового набора данных.

Датасет	Precision	Recall	F1-score
JNLPBA	62.91	59.40	61.10
JNLPBA Wang STM	69.60	74.95	72.17
BC2GM	64.53	61.12	62.78
BC2GM Wang STM	77.50	78.13	77.82

Таб.4 Полученные результаты

В упомянутой выше статье использовалась более глубокая нейронная сеть, которая тренировалась не только на белках, но также на генах, названий болезней и химических элементах. Можно сделать вывод, что лучше тренировать модель на более разнообразных данных для лучшего обобщения и обучения более высокоуровневому шаблона наименований биохимических сущностей.

Модель часто ошибается в комплексных названиях, которые присутствуют в тренировочном наборе данных в небольшом количестве. Такие названия состоящих из нескольких слов, даже при правильном нахождении нескольких частей, модель склонна неправильно классифицировать хотя бы одну часть. Например, если в названии белка есть само слово “белок”, модель видя как часто это слово используется вне контекста имени, почти всегда относит его к обычному слову, а не составляющему имени.

Для решения данной проблемы можно предложить увеличить размер аннотированных данных или ввести такое понятие, как вес вес каждого примера из набора данных. Таким образом можно взвешивать каждый

пример исходя из сложности и длины названия, заставляю таким образом нейронную сеть более тщательно учиться на редких примерах.

Еще одним возможным улучшением является пост-обработка результатов. Т.е. после того, как модель предсказала результаты для тестового набора данных, их можно проанализировать следующим образом: считать расстояние в виде слов между найденными частями белковых имен и основываюсь на заранее согласованных правилах переклассифицировать слова между ними как белки. Похожие подходы используются также без нейронных сетей, т.е. создаются словари известных или часто встречаемых частей названий белков, а далее на основе правил классифицируются слова, находящиеся близко от слов из словаря.

ЗАКЛЮЧЕНИЕ

Таким образом, были изучены поставленные задачи и их важность в практическом применении для извлечения белок-белковых взаимодействий. Также найдены и проанализированы данные для задачи. Построены и реализованы нейронные сети на основе различных архитектур. Проведен сравнительный анализ результатов для различных архитектур.

Помимо этого, полученные результаты для первой поставленной задачи были сравнены с результатами биоинформатического соревнования для аналогичной задачи. При этом разработанная нейронная сеть работает на уровне лучших 5 решений, предложенных на соревновании.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. http://www.ncbi.nlm.nih.gov/entrez/query/static/ overview.html
2. Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, Eisenberg David. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30(1):303–5.
3. Bader Gary D, Betel Doron, Houge Christopher WV. BIND: the biomolecular interaction network database. Nucleic Acids Res 2003;31(1):248–50.
4. Hermjakob Henning, Montecchi-Palazzi Luisa, Lewington Chris, Mudali Sugath. IntAct: an open source molecular interaction database. Nucleic Acids Res 2004;1(32Database issue):452–5.
5. von Mering Christian, Jensen Lars J, Snel1 Berend, Hooper Sean D. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005;33(Database issue):433–7.
6. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In: Proceedings of Pacific symposium biocomputing. 2000. p. 517–28.
7. David PA, Buxton Corney Bernard F, Langdon William B, Jones David T. BioRAT: extracting biological information from full-length papers. Bioinformatics 2004;20(17):3206–13.
8. Rzhetsky Andrey, Iossifov Ivan, Koike Tomohiro, Krauthammer Michael, Kra Pauline. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Informatics 2004;37(1):43–53.
9. Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases. In: Proceedings of the Pacific symposium on biocomputing, Hawaii, USA, 2002. p. 326–37.
10. Pearson H. Biology's name game. Nature 2001;411(6838):631–2.
11. Leser Ulf, Hakenberg Jo¨rg. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 2005;6(4):257–69.
12. Sahu S, Anand A. Recurrent neural network models for disease name recognition using domain invariant features. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume

1: Long Papers). Berlin: Association for Computational Linguistics; 2016. p. 2216–25.

13. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017;33(14):37–48

14. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: HLT-NAACL. San Diego: The Association for Computational Linguistics; 2016. p. 260–70.

15. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan; 2013. p. 39–43

16. Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics; 2009. p. 147–55

17. Wang X, Zhang Y, Ren X, Zhang Y, Zitnik M, Shang J, Langlotz C, Han J. Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics. 2018. ISSN = 1367-4803, https://doi.org/10.1093/

ПРИЛОЖЕНИЕ А

Ссылки на реализацию разработанных алгоритмов:

https://colab.research.google.com/drive/1RwqquZopsaNogz-ErS3kT0uRdlPyYbwe#scrollTo=zGiONeN6bHih

https://colab.research.google.com/drive/1DAjpoise4YeFWk9PfFG0XM8TTgxkGXcf#scrollTo=1WAksiCV0N0_