Федеральное государственное автономное
образовательное учреждение высшего образования
«Казанский (Приволжский) федеральный университет»

На правах рукописи

Мифтахутдинов Зульфат Шайхинурович

Модели связывания именованных сущностей в биомедицинском домене

РЕЗЮМЕ ДИССЕРТАЦИИ
на соискание учёной степени
кандидата компьютерных наук

Казань — 2022

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования "Казанский (Приволжский) федеральный университет".

Научный руководитель: Тутубалина Елена Викторовна, к.ф.-м.н., Казанский (Приволжский) Федеральный Университет.

1 Введение

Тема диссертации

Данная работа посвящена разработке методов нормализации медицинских концептов или связывания именованных сущностей в медицинском домене. Задача связывания именованных сущностей заключается в сопоставлении фразе на естественном языке соответствующего концепта из базы знаний, если таковой имеется. Под концептом в данном случае понимают элемент базы знаний, отражающий некоторое понятие в определенной области знаний. Например, в базе знаний UMLS концепт с идентификатором C0004057 соответствует медицинскому понятию лекарственного средства “Аспирин”. Помимо идентификатора и наименования концепт также может иметь различные синонимы и связи с другими концептами. Таким образом, задача нормализации медицинских концептов состоит в связывании фрагмента текста с конкретным понятием из базы знаний.

Несмотря на то, что задача связывания именованных сущностей является широко исследованной в общем домене, в рамках медицинского домена имеется ряд характерных особенностей: (i) большое многообразие баз знаний, которые зачастую не являются статичными и обновляются с различной периодичностью; (ii) сложность создания соответствующих корпусов с достаточным уровнем покрытия концептов ввиду требования высокой квалификации к аннотаторам; (iii) сильная вариативность форм употребления в рамках одного концепта - одно и то же лекарственное средство может иметь принятые химиками различные названия и множество торговых наименований. В диссертационной работе проведен анализ, модификация и синтез существующих подходов к решению данной задачи. В частности, в работах [1–4] проведена оценка классификационного подхода и предложен вектор семантической близости, характеризующий степень сходства сущности с каждым концептом терминологии. Показаны существенные недостатки в методологиях оценки качества решения задачи нормализации. А именно, показано высокое пересечение (до 60%) между тренировочными и тестовыми частями выборок. Для более реалистичной оценки моделей предложено разбивать выборку на непересекающиеся тренировочные и тестовые части. В перечисленных выше работах показана эффективность предлагаемых векторов семантической

близости и способа внедрения их в классификационный подход. Однако, существенным недостатком классификационных подходов является неспособность распознать концепты отсутствующие в обучающей выборке. В связи с этим, в работах [5; 6] был предложен подход на основе метрического обучения к решению задачи связывания именованных сущностей. Такой подход основан на построении единого векторного пространства для сущностей и концептов. Единое пространство позволяет осуществлять нормализацию на основе мер сходства и рассматривать связывание именованных сущностей как задачу ранжирования. В работе [7] предложен метод объединения классификационного и метрического подходов на основе порогового значения. Оценка предложенного подхода проведена в рамках открытого тестирования Social Media Mining for Health Applications (SMM4H) 2019-го (Task 3), 2020-го (Task 3) и 2021-го (Task 1c) годов [8–10]. По результатам тестирований предложенный подход показал наилучшие результаты среди всех команд участников.

Актуальность работы

Огромный объем текстовых данных в различных источниках представляет широкие возможности для использования их в качестве ресурса для здравоохранения. В качестве рассматриваемых источников данных могут выступать социальные сети, базы научных статей, патентов и клинических испытаний.

Через интернет-ресурсы пользователи получают возможность обмена мнениями и почти неограниченный доступ к информации о сегментах фармацевтического рынка и сведениях медицинской направленности. Кроме того, клинические испытания не всегда позволяют обнаружить полный перечень побочных эффектов. Это вызвано тем, что зачастую побочные эффекты проявляют себя после длительного приема препарата или же оказывают эффект только на определенную группу пациентов, не участвовавшую в клинических испытаниях. Данный факт приводит к наличию большого объема комментариев, содержащих неисследованные побочные эффекты для конкретных лекарственных средств. Использование доступных в сети Интернет комментариев медицинской направленности и выход за рамки простого поиска по ключевым словам представляет собой как возможность, так и сложное и актуальное направление в области обработки текстов на естественном языке.

Применение интеллектуальных методов обработки текста к интернет-ресурсам позволит своевременно выявлять новые побочные эффекты, находить случаи применения лекарств не по назначению, что в свою очередь даст возможность генерировать кандидатов к перепрофилированию лекарственных средств.

Вторым важным ресурсом для здравоохранения являются базы научных статей. Одной из таких баз является PubMED [11], в которой индексируются биомедицинские статьи. По своему содержанию она в большой степени представляет интерес для ученых, занятых исследованиями в области медицины или разработкой новых лекарств. Ключевым моментом при использовании подобного рода баз научных статей является возможность быстрого доступа к нужной информации. Данную проблему можно попытаться решить поисковыми системами общего назначения. Однако возникают сложности с формированием запросов к таким базам данных, поскольку они направлены на получение более точной информации. Например, ученому может потребоваться изучить все работы, в которых проводились совместные исследования определенного гена и заболевания, или работы, в которых гены взаимодействовали определенным образом. Характер таких запросов во многом определяет методы и инструменты построения поисковых систем. В частности, решение задач извлечения и связывания именованных сущностей являются ключевыми в таких поисковых сценариях. Аналогичные утверждения можно сделать и относительно баз медицинских патентов и клинических испытаний.

Таким образом, как в случае работы с интернет ресурсами, так и в случае обработки баз научных статей, патентов и клинических испытаний, одним из важных и зачастую неотъемлемым этапом извлечения структурированной информации из большого объема текстовых данных является связывание именованных сущностей. Безусловно, этому этапу должен предшествовать этап извлечения именованных сущностей, который, однако, в данной диссертационной работе не рассматривается.

Традиционные подходы к нормализации медицинских концептов основываются на использовании словарей и баз знаний. Наиболее распространенной системой сопоставления текстов с концептами из UMLS, основанной на знаниях, является MetaMap [12]. Данная лингвистическая система использует лексический поиск, основанный на генерации различных вариантов употребления нормализуемой фразы. При этом каждому сгенерированному варианту ставится в соответствие некоторая оценка, характеризующая его близость к

исходной фразе. Далее на основе полученных оценок выбирается вариант с максимальной оценкой, имеющий точное совпадение в базе знаний UMLS. Существенным недостатком такого подхода является низкое значение метрики полноты. Следующим подходом, использовавшимся применительно к задаче нормализации медицинских концептов, является обучение ранжированию. Данный подход впервые применен к задаче нормализации в работе [13]. Разработанная авторами система DNorm использует попарное обучение ранжированию, признаками для которого служат векторные представления упоминаний и кандидатов терминов из UMLS. Векторное представление формируется на основе показателей TF-IDF. Описанная в работе [14] система TaggerOne является продолжением работы [13]. Отличие TaggerOne от DNorm заключается в том, что TaggerOne использует марковские и полумарковские модели для совместного обучения задачи извлечения именованных сущностей и нормализации медицинских концептов. В последние годы наблюдается тенденция к рассмотрению задачи нормализации медицинских концептов в большей степени с точки зрения классификационного подхода. К примеру, в работе [15] применены сверточные нейронные сети. Авторы в своей статье показали, что применение моделей глубокого обучения дает значительный прирост по F-мере в сравнении с классическими подходами. В работах, рассматриваемых в рамках данной диссертации, также изучены классификационные подходы, а именно рассмотрены сверточные и рекуррентные архитектуры нейронных сетей на основе векторных представлений и предварительно обученные языковые модели ELMo и BERT; предложены векторы семантической близости и метод их интегрирования в классификационный подход. Предложенные методы показали свою эффективность в рамках открытых тестирований по решению задачи нормализации медицинских концептов CLEF 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020 Task 3. Однако в ходе проводимых исследований были выявлены недостатки классификационного подхода и стандартных методов оценки качества моделей. Одной из серьезнейших проблем классификационного подхода является отсутствие обучающих выборок, покрывающих все возможные медицинские концепты. В связи с указанным ограничением корпусов классификационные подходы не способны распознать концепты, отсутствующие в обучающей выборке. Методы, основанные на правилах, лишены подобного рода ограничений. Однако подходы, основанные на правилах, обладают низкой полнотой извлекаемых

данных. Как следствие, наблюдается необходимость в разработке новых методов решения задачи нормализации медицинских концептов, основанных на современных подходах к обработке естественного языка и не требующих наличия всех медицинских концептов в обучающей выборке. Обучение метрике (metric learning) является одним из таких подходов, так как не требует наличия всех концептов в обучающей выборке и, более того, как было показано в работе [5], является устойчивой к смене словаря. В работе [5], рассматриваемой в рамках диссертации, предложен подход на основе обучения метрике к решению задачи связывания именованных сущностей. Такой подход основан на построении единого векторного пространства для сущностей и концептов. Единое пространство позволяет осуществлять нормализацию на основе мер сходства и рассматривать связывание именованных сущностей как задачу ранжирования. В работе [7] предложен метод объединения классификационного и метрического подходов на основе порогового значения. Оценка предложенного подхода проведена в рамках открытого тестирования SMM4H 2019-го (Task 3), 2020-го (Task 3) и 2021-го (Task 1c) годов. По результатам тестирований предложенный подход показал наилучшие результаты среди всех команд участников. Предложенные в работах [5;6] подходы интегрированы в процессы обработки данных в компании Insilico Medicine. В диссертации описываются некоторые используемые в этой платформы модели и приводятся оценки качества на стандартных наборах данных.

Целью диссертационного исследования является разработка набора актуальных и эффективных средств интеллектуальной обработки информации, решающих задачу связывания именованных сущностей с помощью глубоких нейронных сетей с использованием обучения метрике (metric learning) и негативного сэмплирования (negative sampling).

2 Основные результаты и выводы

Вклад. Основным вкладом работы являются модели связывания именованных сущностей:

1. Модели связывания именованных сущностей, основанные на классификационном подходе. Предложены признаки семантической близости, показавшие свою эффективность в рамках соревнований CLEF eHealth 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020 Task 3,

SMM4H 2021 Task 1c. Показаны недостатки классификационного подхода и стандартных методов оценки качества моделей. В частности, отсутствие обучающих выборок, покрывающих все возможные медицинские концепты, и большое количество объектов тестовой выборки, дублирующих элементы тренировочной выборки. Предложен метод оценки моделей, устраняющий высокий уровень пересечения обучающих и тестовых выборок.

2. Модель связывания именованных сущностей, основанная на метрическом подходе. Показана устойчивость данной модели к смене словаря и способность распознавать концепты, не присутствовавшие в обучающей выборке.
3. Модель связывания именованных сущностей, основанная на комбинированнии классификационного и метрического подходов. Показана эффективность данного метода в рамках соревнований SMM4H 2020 Task 3 и SMM4H 2021 Task 1c.

Теоретическая и практическая значимость

Практическая значимость результатов определяется тем, что разрабатываемые модели направлены на анализ текстов из открытых источников, в том числе из сети Интернет, где содержится обширный набор информации медицинской направленности, который может быть использован для улучшения здравоохранения и в исследовательской деятельности специалистов. Теоретическая значимость заключается в предложеных в диссертационной работе новых моделях связывания именованных сущностей. В первую очередь улучшены модели, основанные на классификационном подходе, и показаны недостатки таких методов и методологии их оценок. Предложены более реалистичные методы к оценке нормализации медицинских концептов. Для решения проблемы ограниченности словаря тренировочных данных предложен метод, основанный на метрическом подходе. Наконец, предложен комбинированный подход, позволяющий объединить сильные стороны обоих методов решения – классификационного и метрического.

Результаты, выносимые на защиту

1. Разработана модель связывания именованных сущностей, основанная на классификационном подходе, с использованием признаков семантической близости.
2. Разработана модель связывания именованных сущностей, основанная на метрическом подходе.
3. Разработана модель связывания именованных сущностей, основанная на комбинированном подходе.

Личный вклад в результаты, выносимые на защиту

В первой статье автором предложены векторы семнтической близости и модели интегрирующие предложенные векторы в классификационный подход. Все эксперименты проведены автором. Во второй и третьей статье автору принадлежат модели, построенные на основе метрического подхода с использованием триплетной целевой функцией. Все эксперименты в указанных статьях проведены автором.

Публикации и апробация работы

Автор диссертации является основным автором в 8 основных статьях по теме диссертации.

Публикации повышенного уровня

1. **Miftahutdinov Z.** et al. Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer (Обучение представлений биомедицинских сущностей лекарств и заболеваний с помощью сети Transformer) //European Conference on Information Retrieval. – Springer, Cham, 2021. [Scopus, ECIR - Core A conf.]
2. **Miftahutdinov Z.**, Kadurin A., Kudrin R., Tutubalina E. Medical concept normalization in clinical trials with drug and disease representation learning (Нормализация медицинских концептов в текстах клинических испытаний посредством обучение представлений биомедицинских сущностей лекарств и заболеваний) //Bioinformatics. – 2021. – Т. 37. – №. 21. – С. 3856-3864 DOI: 10.1093/bioinformatics/btab474 (Q1, Impact Factor 2021 6.64) [Scopus]

3. Tutubalina, E., **Miftahutdinov, Z.**, Nikolenko, S., & Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks (Нормализация медицинских сущностей из социальных медиа-ресурсов с использованиесм рекуррентных нейронных сетей). Journal of biomedical informatics. — Vol. 84. — Pp. 93–102 DOI:10.1016/j.jbi.2018.06.006 (Q1, Impact Factor 2019 3.5) [Scopus, WOS]

Доклады на конференциях

1. The 10th International Conference on Analysis of Images, Social Networks and Texts, December 16, 2021, keynote. Тема: "Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer".
2. European Conference on Information retrieval, March 28, 2021. Тема: "Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer".
3. European Conference on Information retrieval, April 14, 2020. Тема: "On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts".
4. The 28th International Conference on Computational Linguistic, December 8, 2020. Тема: "Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models".
5. The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, July 28, 2019. Тема: "Deep Neural Models for Medical Concept Normalization in User-Generated Texts".

3 Содержание работы

3.1 Классификационный подход и предложенные признаки семантической близости

Одним из устоявшихся в последнее время подходов к нормализации медицинских концептов является классификационный. Современные алгоритмы классификации базируются на глубоких нейронных сетях с функцией активации softmax на последнем слое. Softmax обобщает логистическую

функцию для многомерного случая. В таком варианте модель обучается моделировать вероятностное распределение для сущности m на множестве всех концептов. Преимуществом такого подхода является высокая точность распознавания концептов присутствовавших в обучающей выборке. Однако одной из серьезнейших проблем классификационного подхода является отсутствие обучающих выборок покрывающих все возможные медицинские концепты. Как было показано в работе [1], корпус CADEC [16] содержит менее 5% концептов медицинской номенклатуры SNOMED-CT. Подобные утверждения верны и для большинства других корпусов. К примеру, корпус PsyTAR [17] содержит менее 1% от исходной системы кодирования, корпус SMM4H – менее 0.5% [18], корпус TwiMed – менее 0.5% [19], корпус NCBI [20] – менее 0.01%, корпус TwADR-L – менее 0.01% [15]. Для обхода подобных ограничений в работе [7], рассматриваемой в данной диссертации, предложен комбинированный подход.

Рассмотрение методов классификации в рамках задачи нормализации медицинских концептов начато автором диссертационной работы в статье [1] и продолжено в работе [2]. Предложенный в работах подход использует два различных способа представления текста на естественном языке в векторном виде. Первый основан на глубоких нейронных сетях, а именно на представлениях, полученных из предобученной языковой модели BioBERT. Второй способ представления основан на семантической близости между исходным выражением и медицинскими понятиями из онтологии UMLS. При этом семантическая близость определяется как косинусное расстояние между векторными представлениями на основе TF-IDF или word2vec медицинских концептов и исходным выражением. Таким образом, каждый элемент полученного вектора показывает степень близости выражения на естественном языке к концептам из медицинской онтологии UMLS. На основе полученных векторных представлений производится классификация к соответствующему медицинскому термину. Описанный подход наглядно представлен на рисунке 1.

Предложенные в статьях [1; 2] архитектуры имеют общую последовательность обработки. На первом этапе происходит преобразование исходного текста в векторное представление с использованием глубоких нейронных сетей согласно формуле ниже:

$$y_m = red(Encoder(m)), \tag{1}$$

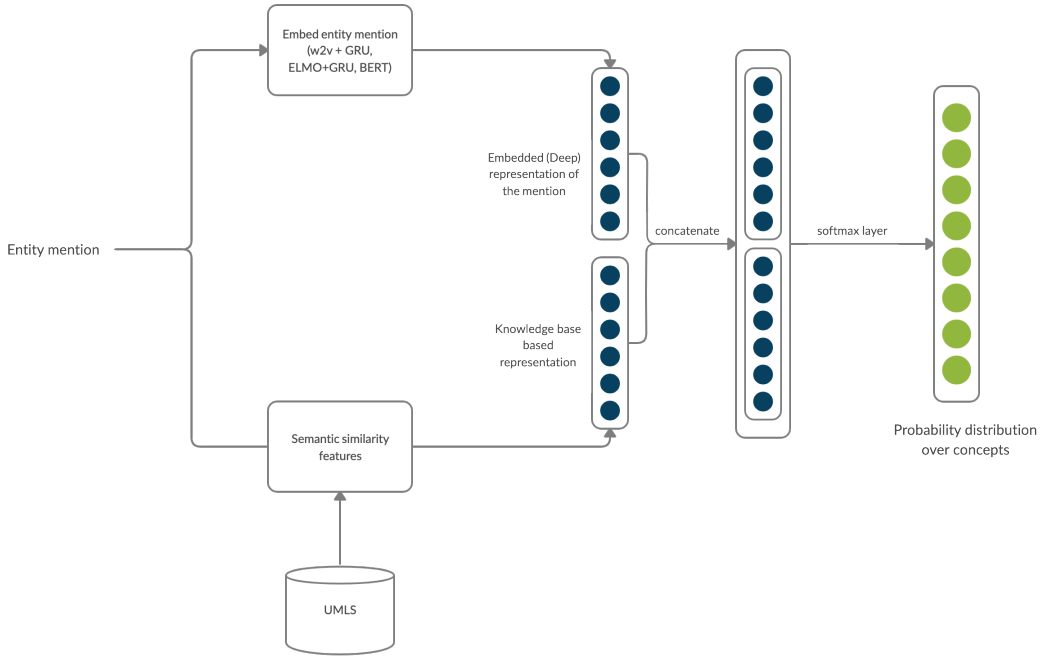

Рис. 1 — Общая схема классификационного подхода

где Encoder – некоторая обучаемая глубокая нейронная сеть, $red(\cdot)$ – функция, которая сводит последовательность векторов в один вектор. В качестве кодирующей сети в работе рассмотрены:

– сверточная нейронная сеть, использующая векторные представления слов
– рекурентная нейронная сеть GRU с векторными представлениями слов
– рекурентная нейронная сеть GRU с векторными представлениями слов, полученными из предобученной языковой модели ELMo [21]
– предобученная языковая модель BERT.

Для сверточных и рекуррентных моделей векторные представления слов инициализировались предобученными векторами word2vec из статьи [22].

Существует несколько различных вариаций функции $red(\cdot)$. В частности, в работах [1; 2] использованы следующие варианты: выбрать выход, соответствующий токену CLS для кодирующей сети BERT; применить механизм внимания представленный в статье [23] для реккурентной кодирующей сети.

Далее к полученному векторному представлению на основе глубоких нейронных сетей конкатенируется вектор семантической близости между

исходным выражением и медицинскими понятиями из онтологии UMLS. На последнем этапе на основе полученных векторных представлений производится классификация к соответствующему медицинскому термину с использованием функции $softmax$, как показано в формуле 2:

$$c_p = softmax([y_m : y_s]), \tag{2}$$

где y_m соотвествует представлению, полученному из глубокой нейронной сети, y_s – вектор семантической близости, $[y_m : y_s]$ – конкатенация обозначенных представлений сущности.

Также в работе [1] предложен подход к оценке моделей. При детальном рассмотрении набора данных CADEC обнаружено сильное пересечение между обучающей и тестовой выборками - а именно, тестовая часть выборки на 60% состояла из примеров тренировочной части. При таком соотношении трудно прогнозировать качество модели на новых данных, так как модель, запоминающая обучающую выборку уже достигает 60% точности. Для более реалистичной оценки моделей решено разбивать выборку на непересекающиеся тренировочные и тестовые части. С этой целью из исходной выборки удалялись дубликаты, далее все сущности группировались в рамках одного концепта. Таким образом получилось n груп, где n – количество концептов, представленных в выборке. Внутри каждой выборки происходило разбиение на тестовую и тренировочную части. Далее все тренировочные и тестовые части объединялись. Данное разбиение гарантирует, что в тестовой части не будет пар (сущность, концепт), присутствующих в обучающей выборке, а также тот факт, что все концепты будут представлены как в тестовой части, так и в тренировочной.

В таблицах 2 и 1 содержатся результаты, представленные в работах [1] и [2] соотвественно. При этом в таблице 1, помимо результатов на непересекающихся подмножествах, приведены результаты на случайном разбиении на тренировочную и тестовую выборки. Как можно заметить, сверточные нейронные сети достигают наименьшего качества среди всех моделей, в 46% точности на корпусе CADEC. Рекуррентные нейронные сети достигают 64.5% точности. Рекуррентные нейронные сети с использованием механизма внимания и векторных представлений word2vec достигают 70,05% точности в случае использования вектора семантических признаков. Модель, не использующая семантические признаки, достигает 66,56% точности. Рекуррентная сеть на основе контекстных векторных представлений слов ELMo на корпусе

Таблица 1 — Метрики качества классификационных моделей на корпусах CADEC, PsyTAR, SMM4H 2019 Task 3. Приведены результаты для случайного (Random) и непересекающегося (Custom) разбиения выборки на тестовую и тренировочные части.

Метод	**CADEC**		**PsyTAR**		**SMM4H 2019 Task 3**
	Random	Custom	Random	Custom	Official
Baseline: match with training set annotation	66.09	0.0	56.04	2.63	67.12
DNorm [15]	73.39	-	-	-	-
CNN [15]	81.41	-	-	-	-
RNN [15]	79.98	-	-	-	-
Attentional Char-CNN [24]	84.65	-	-	-	-
Hierarchical Char-CNN [25]	-	-	-	-	87.7
Ensemble [26]	-	-	-	-	88.7
GRU+Attention	82.19	66.56	73.12	65.98	83.16
GRU+Attention w/ TF-IDF (MAX)	84.23	70.05	75.53	68.59	86.28
ELMo+GRU+Attention	85.06	71.68	77.58	68.34	86.60
ELMo+GRU+Attention w/ TF-IDF (MAX)	85.71	74.70	79.52	70.05	87.52
BERT	88.69	79.83	83.07	77.52	89.28
BERT w/ TF-IDF (MAX)	88.84	79.25	82.37	77.33	89.64

CADEC имеет 71,68% точности в варианте без использования семантических признаков и 74,70% точности – с использованием семантических признаков. Модель архитектуры BERT достигает 79,83% с семантическими признаками и 79,25% – без обозначенных признаков. Аналогичные результаты получаются и на корпусах PsyTAR и SMM4H 2019 Task 3. Как видно из результатов, контекстные векторные представления ELMo дают лучшие показатели качества в сравнении с неконтекстными представлениями word2vec. Однако наилучшие показатели качества достигаются при использовании модели BERT. Также следует отметить, что вектор семантических признаков дает прирост только в случае использования векторных представлений word2vec и ELMo. Приведенные выше наблюдения аналогичны для всех корпусов, использованных в статье [2].

Оценка предложенной модели на основе контекстных представлений BERT также проведена в рамках открытого тестирования SMM4H 2019 Task 3. Задача, решаемая в рамках соревнования SMM4H 2019 Task 3, состоит из двух частей: извлечение именованных сущностей и нормализация медицинских концептов. Для извлечения именованных сущностей использовался подход на основе языковой модели BERT с классификационным слоем для каждого токена по BIO схеме. В таблице 3 приведены результаты как для компоненты извлечения именованных сущностей, так и для комплексной оценки предложенного решения. Как видно из таблицы, классификатор, построенный на

Таблица 2 — Метрики качества классификационных моделей на корпусе CADEC. Приведены результаты для непересекающегося разбиения выборки на тестовую и тренировочные части.

Метод	**Parameters**	**Accuracy**
CNN	HealthVec, 100 feature maps	46.19
CNN	PubMedVec, 100 feature maps	45.79
LSTM	HealthVec, 200 hidden units	64.51
LSTM	PubMedVec, 200 hidden units	64.24
GRU	HealthVec, 200 hidden units	63.05
GRU	PubMedVec, 200 hidden units	62.73
LSTM+Attention	HealthVec, 200 hidden units	65.73
LSTM+Attention	PubMedVec, 200 hidden units	64.92
LSTM+Attention	HealthVec, 100 hidden units	64.83
GRU+Attention	HealthVec, 200 hidden units	**67.08**
GRU+Attention	PubMedVec, 200 hidden units	66.55
GRU+Attention	HealthVec, 100 hidden units	66.56
with prior knowledge		
LSTM+Attention	HealthVec, 100, similarity: TF-IDF (ALL)	67.63
LSTM+Attention	HealthVec, 200, similarity: TF-IDF (ALL)	66.83
GRU+Attention	HealthVec, 100, similarity: TF-IDF (ALL)	69.92
GRU+Attention	HealthVec, 200, similarity: TF-IDF (ALL)	69.42
GRU+Attention	HealthVec, 100, similarity: W2V (ALL)	69.14
GRU+Attention	HealthVec, 100, similarity: TF-IDF (MAX)	**70.05**

предобученной языковой модели BERT с использованием признаков семантической близости, показал наилучшие результаты среди остальных участников.

Предложенные в вышеописанных работах признаки семантической близости также применены в другой вариации нормализации медицинских концептов. В некоторых случаях задача нормализации медицинских концептов требует сопоставления фразы с несколькими концептами из медицинской базы знаний. Такая вариация часто встречается при обработке клинических текстов и требует приведения сущностей к Международной Классификации Болезней (МКБ). Международная классификация болезней – диагностическая система, которая используется для мониторинга и классификации

Таблица 3 — Метрики качества классификационной модели на корпусе SMM4H 2019 Task 3. Результаты других подходов взяты из статьи [8]. EF_1 – F-мера частичного совпадения для подзадачи извлечения именованных сущностей, NF_1, NP, NR – соответственно F-мера, точность и полнота комплексной оценки предложенных решений.

Метод	**E F_1**	**N F_1**	**N P**	**N R**
KFU@NLP Team	**66.0**	**43.0**	**36.0**	**54.0**
ensemble RNN & Few-Shot Learning	-	35.0	34.0	36.0
BERT + Flair + RNN	63.0	31.0	37.0	27.0
encoder-decoder (W biLSTM + attention)	60.0	21.0	22.0	20.0

проблем со здоровьем и смерти, а также для предоставления информации в клинических целях.

Для решения обозначенной задачи ранее предлагалось использовать методы, основанные на правилах и словарях. В частности, в работе [27] применено решение с использованием системы полнотекстового поиска Solr. Описанная авторами система достигла 84.8% точности на наборе данных [28]. Cabot et al. в своей работе [29] применили комбинацию подхода на основе словаря и алгоритмов нечеткого соответствия. Их система достигла 80,38% точности. Наряду с подходами, основанными на правилах и словарях, использовался классификационный подход. В частности, в работе [30] использовалась многозначная классификация на основе метода опорных векторов (SVM) совместно с этапом предварительной обработки текста (удаление стоп-слов, удаление диакритических знаков, исправление некоторых орфографических ошибок). Предложенный авторами метод достиг 84.7% F-меры.

В статьях [3; 4], рассматриваемых в данной диссертации, автором предложен новый метод установления соответствия между текстами из медицинских документов и формальными медицинскими понятиями, основанный на глубоких нейронных сетях архитектуры “кодировщик–декодировщик” (encoder-decoder). В данном случае задача нормализации рассмотрена как задача перевода последовательности токенов в последовательность кодов из МКБ соответствующих исходной фразе (сущности).

Для реализации модели "кодировщик-декодировщик"использована архитектура, представленная в работе [31]. В качестве кодирующей сети реализована двунаправленная LSTM сеть. В качестве декодирующей сети применена однонаправленная LSTM сеть.

В предложенный метод интегрированы словарные признаки, основанные на знаниях о медицинских терминах и соответствующих кодах в международной классификацией болезней (МКБ-10, ICD-10). Реализованная система установления соответствия между текстами из сертификатов о смерти на английском языке и МКБ-10 получила наилучший результат в одной из дорожек по итогам открытого тестирования систем CLEF eHealth 2017 Task 1.

3.2 Обучение метрике и негативное сэмплирование

Задачу нормализации медицинских концептов можно рассматривать с точки зрения информационного поиска. В данном случае сущность, которую необходимо привести к концепту из медицинской базы знаний, можно интерпретировать как запрос, множество концептов – как коллекцию документов, которые необходимо упорядочить по мере убывания релевантности. При этом первым элементом в ранжированном списке должен находиться концепт, соответствующий сущности. В такой постановке документ состоит из полей, представляющих концепт: уникального идентификатора концепта и множества наименований концепта. Помимо перечисленных двух полей, различные медицинские базы знаний могут содержать дополнительную информацию. К примеру, иерархическую структуру концептов или семантический тип концепта [32]. Однако, в данной работе основными полями для осуществления нормализации медицинских концептов являются уникальный идентификатор и множество наименований.

Данный подход впервые применен к задаче нормализации в работе [13]. В разработанной авторами системе DNorm применяется попарное обучение ранжированию, в котором в качестве признаков задействованы векторные представления упоминаний и кандидатов терминов из UMLS. Векторы представлений формируются как TF-IDF представления. Описанная в работе [14] система TaggerOne является продолжением работы [13]. Отличие TaggerOne от DNorm заключается в том, что TaggerOne использует марковские и полумарковские модели для совместного обучения задаче извлечения именованных сущностей и нормализации медицинских концептов. Система достигает 82.9% F-меры на корпусе NCBI. Однако, данные работы основаны на устаревших n-граммных представлениях текста, не предоставляющих широких возможностей современных подходов обработки текста.

В статье [5], рассматриваемой в данной диссертации, предложен подход к тонкой настройке языковой модели на основе архитектуры трансформер с использованием метрического обучения (metric learning), в частности, триплетной целевой функции (triplet loss), и негативного сэмплирования (negative sampling). Модель, полученная в результате обучения метрическим подходом на медицинских данных, далее именуется как DILBERT. Обозначенный подход позволяет строить общее семантическое пространство векторов для сущностей и концептов из базы знаний, в котором схожие по смыслу тексты находятся близко друг к другу. Данное свойство позволяет ранжировать концепты на основе некоторой функции расстояния s и решать задачу нормализации медицинских концептов.

Следуя обозначениям, предложенным в [33], как сущности, так и концепты преобразуются в векторные представления следующим образом:

$$y_m = red(T(m)); y_c = red(T(c)), \tag{3}$$

где T – глубокая нейронная сеть архитектуры трансформер, веса которой могут обновляться во время тонкой настройки, $red(\cdot)$ – функция, которая редуцирует последовательность векторов в один вектор, m – сущность, которую необходимо привести к соотвествующему концепту, c – наименования концепта. Существует несколько различных реализаций функции $red(\cdot)$, в том числе: выбрать выход, соответствующий токену CLS или вычислить поэлементное среднее по всем векторам, чтобы получить вектор фиксированного размера. Эмпирически установлено, что усреднение является оптимальной реализацией функции $red(\cdot)$. В качестве предварительно обученной модели архитектуры трансформер используется модель BioBERT v1.1., предварительно обученная на корпусе аннотаций биомедицинских статей PubMED.

Оценка релевантности кандидата c_i для сущности m задается функцией расстояния, примененной к соотвествующим векторным представлениям. В диссертационной работе рассмотрены Евклидово и косинусное расстояния. Однако в ходе проведения экспериментов не установлено значимого отличия между данными двумя вариантами по результирующим метрикам качества. В связи с чем метрики качества приводятся только для Евклидова расстояния.

$$s(m, c_i) = ||y_m - y_{c_i}||, \tag{4}$$

Как показано в работе [34], можно осуществить тонкую настройку (fine-tuning) языковой модели BERT для того, чтобы векторные представления,

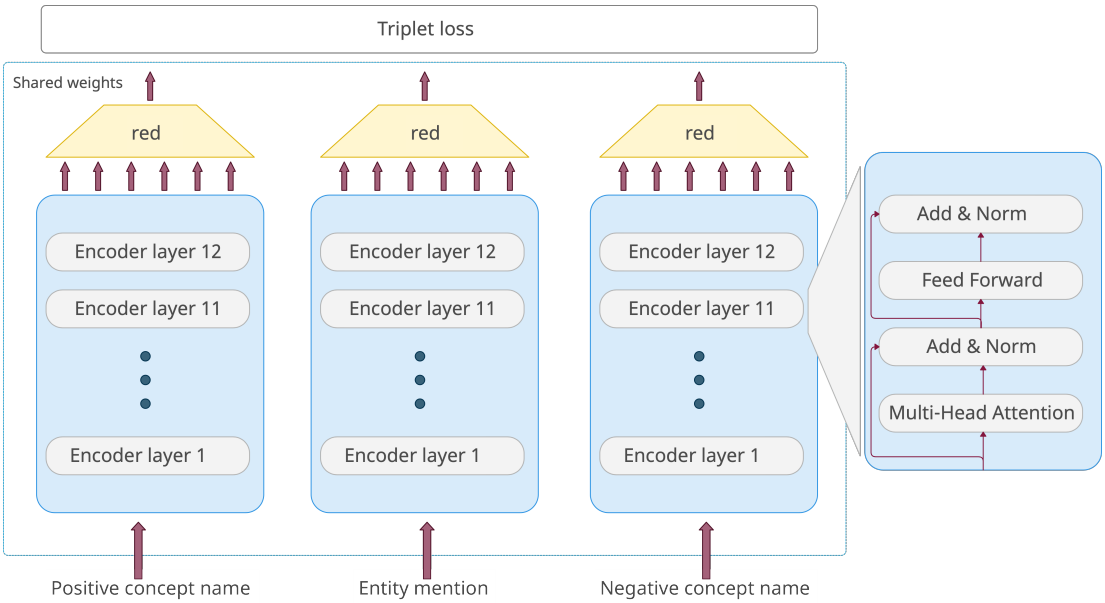

Рис. 2 — Архитекутра DILBERT. Модель состоит из трех кодирующих сетей архитектуры Transformer с общими весами. Сеть слева кодирует положительный пример, сеть посередине – сущность, правая – негативный пример. Также на рисунке отдельно продемонстрирован блок, из которого состоят кодирующие сети.

полученные из этой модели, более точно выражали семантическую близость. Тонкая настройка модели позволяет располагать сущности и соотвествующие концепты в векторном пространстве ближе и отдалять нерелевантные концепты. Для приближения векторных представлений сеть обучается с использованием триплетной целевой функции. Пусть дано упоминание сущности m, наименование позитивного концепта c_g, наименование негативного концепта c_n, триплетная целевая функция настраивает нейронную сеть таким образом, чтобы расстояние между m и c_g было меньше, чем расстояние между m и c_n. Математически минимизируется следующая функция потерь:

$$max(s(m, c_g) - s(m, c_n) + \epsilon, 0), \tag{5}$$

где ϵ – отступ, который гарантирует, что c_g по крайней мере на ϵ ближе к m, чем к c_n. В проведенных экспериментах $\epsilon = 1$. Схематически модель проиллюстрирована на рисунке 2.

Подбор позитивных и негативных примеров является существенно важной компонентой для триплетной целевой функции. Рассмотрим далее

процедуру генерации для каждой из них. Предположим, что дана пара: сущность с соответствующим ей идентификатором концепта, а также словарь. Для генерации положительных примеров словарь ограничен концептами, которые имеют тот же идентификатор, что и сущность. При генерации негативных примеров [35] используется остальная часть словаря. Для выбора положительных и отрицательных примеров исследуется несколько стратегий:

- **Случайная генерация** (random sampling): позитивные и негативные примеры случайным образом выбираются из соотвествующих частей словаря;
- **Случайная генерация с учетом иерархии** (random sampling + n parents): к случайно сгенерированным позитивным примерам добавляются наименования родительских концептов. Негативные примеры генерируются также случайным образом.
- **Повторная генерация** (resampling): в данном случае генерация происходит с использованием модели, обученной на случайным образом сгенерированных триплетах. Генерация осуществляется в несколько этапов: (i) кодируются все упоминания и наименования концептов с использованием текущей модели, (ii) выбираются положительные примеры с тем же идентификатором концепта, которые наиболее близки к упоминанию сущности, (iii) выбираются негативные примеры, которые наиболее близки к упоминанию сущности и при этом имеют отличный идентификатор концепта.
- **Повторная генерация с учетом иерархии** (resampling + n siblings): позитивные и негативные примеры генерируются аналогично предыдущему пункту. Однако дополнительно в качестве негативных примеров выбираются наименования концептов, имеющие общего родителя с концептом, соотвествующим сущности.

Оценка предложенного метода проводилась на корпусе CDR. Для оценки качества применена наиболее часто используемая в области нормализации медицинских концептов метрика - точность (Accuracy). Как было показано в [36], корпус CDR содержит большое количество дубликатов и сильное пересечение между обучающими и проверочными подвыборками в рамках одного корпуса. Чтобы исключить влияние способности модели к запоминанию тренировочной выборки и получить более реалистичные метрики качества, оценка производилась на refined тестировочной выборке, представленной в работе [36]. В качестве базовой модели для сравнения использовалась языковая

Таблица 4 — Показатели точности (accuracy) для модели DILBERT на корпусах CDR Disease и CDR Chemicals.

Метод	**CDR Disease**	**CDR Chemical**
BioBERT ranking	66.4	80.7
BioSyn	74.1	**83.8**
DILBERT, random sampling	75.5	81.4
DILBERT, random + 2 parents	75.0	81.2
DILBERT, random + 5 parents	73.5	81.4
DILBERT, resampling	**75.8**	83.3
DILBERT, resampling + 5 siblings	75.3	82.1

модель BioBERT v1.1. Также качество модели сравнивалось с наиболее эффективной на момент проведения исследования моделью BioSyn [37]. Результаты приведены в таблице 4.

Одной из ключевых особенностей метрического подхода является возможность определения сущностей, для которых не существует подходящего концепта в словаре. При этом методология естественным образом вытекает из предположения о том, что схожие по значению элементы находятся близко друг к другу. Соответственно, если все объекты словаря находятся достаточно далеко от нормализуемой сущности, то данная сущность не имеет подходящего концепта. Данное утверждение формализуется пороговым значением. Таким образом, если все концепты находятся на отдалении большем, чем пороговое значение t, можно заключить, что сущности не соответствует ни один из концептов. Для определения порога используются максимальное расстояние истинно положительных примеров (True Positive cases) d_{tp} и минимальное расстояние ложно положительных примеров (False Positive cases) d_{fp}. Пороговое значение устанавливается равным взвешенной сумме:

$$t = a_1 * d_{tp} + a_2 * d_{fp}, \tag{6}$$

где a_1 – доля истинно положительных примеров среди сущностей, ближайший концепт для которых находится на расстоянии $s \in [d_{fp}; d_{tp}]$; a_2 – доля ложно положительных примеров в том же множестве сущностей. При этом, если множество сущностей, по которым определяются веса a_1 и a_2, пусто, то коэффициенты устанавливаются равными $\frac{1}{2}$. Оценка предложенного подхода определения сущностей, не имеющих соответствующего концепта в словаре,

Таблица 5 — Качество модели DILBERT на корпусе клинических испытаний. Результаты представлены для типов сущности заболевание (CT Condition) и лекарства (CT Intervention). При этом показаны метрики качества как для подкорпуса состоящего только из сущностей с единственным концептом (single concept) так и для всего корпуса (full set)

Model	**CT Condition**		**CT Intervention**	
	single concept	full set	single concept	full set
BioBERT ranking	72.60	71.74	77.83	56.97
BioSyn	86.36	-	79.58	-
DILBERT with different sampling strategies				
random sampling	85.73	84.85	**82.54**	**81.16**
random + 2 parents	86.74	86.36	81.84	79.14
random + 5 parents	**87.12**	**86.74**	81.67	79.14
resampling	85.22	84.63	81.67	80.21
resampling + 5 siblings	84.84	84.26	80.62	76.16

проведена на корпусе клинических испытаний [5]. Результаты приведены в таблице 5. Еще одной ключевой особенностью предложенного подхода является независимость от терминологии, представленной в обучающей выборке. Модель, полученную на одной терминологии, можно без дополнительного обучения применить на другой. В частности, в таблице 5 приведены результаты для моделей обученных на корпусе CDR, привязанной к терминологиям MEDIC и CTD, и оценка которых осуществлена на корпусе с внутрикорпоративной терминологией компании InSilico и MeSH терминологией. Модель достигла 87.12% и 90.53% точности (accuracy) на внутрикорпоративном и MeSH словарях соответственно.

Как видно из таблиц 4 и 5, модель DILBERT показывает результаты на уровне с моделью BioSyn на корпусе CDR Chemical и превосходит на корпусах CDR Disease, CT Condition, CT Intervention.

Рассмотренные модели вычисляют сходство между сущностями и наименованиями концептов на уровне слов и частей слова. Данная особенность позволяет связать сущности, которые похожи на наименования концептов из терминологии, такие как «дефекты зрения» и «нарушение зрения». Однако эти модели могут неправильно связывать сущности с концептами из базы знаний в двух основных случаях: (i) форма употребления сущности

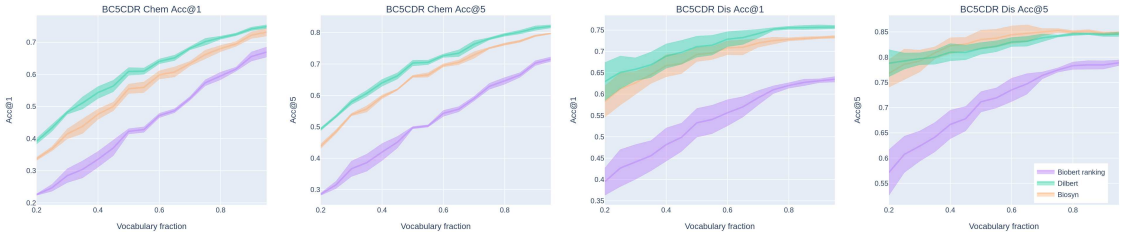

Рис. 3 — Оценка влияния полноты словаря на качество модели.

и наименование концепта схожи, но имеют различные значения (например, “хлорфенак” (C041190) и “хлорферон” (C305311))), (ii) оба выражения имеют одно и то же значение, но различаются по форме употребления (“метиндол” и “индометацин” являются одним и тем же противовоспалительным препаратом (D007213)). Более того, базы знаний могут быть устаревшими, а охват синонимов в них может быть неполным.

Проведена серия экспериментов с моделями, обученными на корпусах CDR Disease & Chemical, с использованием неполного словаря во время прогнозирования. Эти эксперименты направлены на проверку того, насколько хорошо модель запоминает наименования концептов из обучающего словаря. Чтобы создать неполные словари, исходные версии словарей группировались по идентификатору концепта, в каждой группе случайным образом генерировалась подвыборка размерности Vocabulary fraction $\times$ group size,где group size – размер группы соответствующего концепта, а Vocabulary fraction – доля синонимов, остающихся в неполном словаре. Отметим, что если количество наименований концептов после выборки оказывалось дробным, то производилось округление количества имен концептов до наименьшего целого числа. Например, 95% из 10 наименований концепта - 9. Доля синонимов варьируется от 0.95 до 0.20 с шагом -0.05. Для каждого значения Vocabulary fraction процедура случайной генерации неполного словаря проводилась четыре раза, далее результаты усреднялись. Согласно результатам, представленным на рисунке 3, можно сделать следующие выводы.

Во-первых, результаты по метрике Acc@1 демонстрируют, что ухудшение показателей от полного словаря до словаря содержащего 30% синонимов является существенным. Данное явление объясняется тем, что модели при подборе концепта в большей степени опираются на наименование, наиболее близкое по форме употребления. Однако, ввиду неполноты словаря, необходимые синонимы могут отсутствовать во время прогнозирования. Также

можно заметить, что эффективность моделей нормализации лекарств снизилась больше, чем моделей нормализации болезней, поскольку названия лекарств очень разнородны (есть названия активных соединений, торговые марки, патентованные идентификаторы и т.д.). Во-вторых, с точки зрения Acc@5, модели показывают не столь значительное падение метрики качества. Наконец, как видно из графиков, модель DILBERT показывает наименьшее падение качества при использовании неполного словаря, по сравнению с другими моделями.

Следует отметить, что предложенный подход позволяет заранее кэшировать и индексировать векторные представления наименований концептов. Что в данном случае существенно ускоряет время, затрачиваемое на обработку одной сущности. Так, при использовании библиотеки FAISS с поддержкой GPU для быстрого поиска в многомерном пространстве на обработку 10 миллионов сущностей затрачивается 3 часа, или порядка 1000 сущностей в секунду. Обработка осуществляется на одной видеокарте Nvidia TITAN X.

3.3 Комбинированный подход

Как было отмечено, классификационные подходы имеют высокую точность распознавания концептов, присутствовавших в обучающей выборке, и не способны распознать остальные концепты. Подходы, основанные на обучении метрик, являются в некотором роде дополнением к классификационным подходам. Так как позволяют распознавать часть концептов, находящихся за пределами обучающей выборки. Вследствие чего автором в работе 2020 года [7] предложен метод к комбинированию классификационного и метрического подходов. При этом нормализация осуществлялась с использованием предсказаний обоих предикторов на основе порогового значения t. Пусть, согласно метрическому подходу, c_m является ближайшим к сущности m концептом, отдаленным на расстояние $s(m, c_k)$ – расстояние определено согласно формуле 4. Концепт c_c является наиболее вероятным в рамках классификационной модели. Тогда выход комбинированной модели определяется по следующей формуле:

$$c_{cm} = \begin{cases} c_m, \text{если } s(m, c_k) < t \\ c_c, \text{в противном случае} \end{cases} \tag{7}$$

По формуле 7 выходом комбинированного подхода будет концепт подобранный метрическим подходом, если его синоним находится на расстоянии не больше, чем t от сущности. В противном случае выходом комбинированной модели будет наиболее вероятный концепт согласно классификационной модели. Таким образом, предпочтение отдается концептам, полученным метрическим подходом и имеющим высокий рейтинг, а затем наиболее вероятным по классификационному подходу. Максимальное расстояние в данном случае является гиперпараметром и подбирается на валидационном наборе данных.

Таблица 6 — Метрики качества комбинированной модели на корпусе SMM4H 2020 Task 3. Результаты других подходов взяты из статьи [9]

Метод	**E F_1**	**N F_1**	**N P**	**N R**
KFU@NLP Team	**76.0**	**46.0**	**48.0**	**45.0**
BERT, CADEC, SMM4H'17 corpus	73.0	38.0	34.0	44.0
RoBERTa, multi-task learning	69.0	35.0	33.0	38.0
BERT ensemble, fastText-based similarity metrics, CADEC	58.0	22.0	24.0	20.0
BiLSTM, CRF, GloVe and EXT word embeddings, QuickUMLS	46.0	20.0	35.0	14.0
-	56.0	15.0	15.0	14.0
dictionary	16.0	0.0	0.0	0.0

Таблица 7 — Метрики качества комбинированной модели на корпусе SMM4H 2021 Task 1c. Результаты других подходов взяты из статьи [10]

Метод	**E F_1**	**N F_1**	**N P**	**N R**
KFU@NLP Team	40.0	**29.0**	30.1	27.5
BERT with joint NER and Normalization	37.0	24.0	37.1	17.8
RoBERTa, multi-task learning	69.0	35.0	33.0	38.0
BERTweet and similarity measures	42.0	20.0	13.9	34.2
Multi-task learning with selective oversampling	51.0	16.0	16.0	17.0

Оценка предложенного метода производилась в рамках открытых тестирований Social Media Mining for Health Applications (#SMM4H) Shared Tasks 2020 Task 3 и Social Media Mining for Health Applications (#SMM4H) Shared Tasks 2021 Task 1c. В контексте SMM4H 2020 и SMM4H 2019 решалась общая

задача, включающая в себя две подзадачи: извлечение именованных сущностей и нормализация медицинских концептов. Для извлечения именованных сущностей использовался подход на основе языковой модели EnDR-BERT [7] с классификационным слоем для каждого токена по BIO схеме. Для решения задачи нормализации использовался вышеописанный комбинированный подход. В качестве языковой модели использовались BERT, BioBERT, SciBERT. Наилучшие результаты для задачи нормализации показал BERT, обойдя другие варианты на 1-2% по показателю точности (accuracy). На валидационном наборе данных SMM4H 2020 предложенный подход показал следующие результаты: показатель F-меры точного совпадения достиг 57.81% при извлечении именованных сущностей, точность (accuracy) при решении задачи нормализации равнялась 45.17%. Оценка на тестовом наборе данных осуществлялась авторами задачи двумя способами: отдельно для распознавания сущностей и в общем для всей задачи. Описанный подход показал наилучшие результаты в рамках соревнования SMM4H 2020 Task 3 и SMM4H Task 1c. На тестовом наборе SMM4H 2020 для извлечения именованных сущностей предложенный подход показал наилучшие результаты, достигнув 75.5% F-меры частичного совпадения. При этом средний показатель среди всех команд составил 56.4% F-меры. Для комплексной оценки описанный метод показал 46.3% F-меры на тестовом наборе данных при среднем равном 29.2%. Подробные результаты на тестовом наборе данных приведены в таблице 7. На тестовом наборе SMM4H 2021 достиг 40% F-меры по точному совпадению для задачи извлечения именованных сущностей, 29% F-меры, 30% точности и 27.5% полноты для комплексной оценки. Стоит отметить, что несмотря на невысокие показатели компоненты извлечения именованных сущностей в открытом тестировании 2021го года, предложенное решение достигло наивысших показателей среди всех участников в комплексной оценке системы. Полученные результаты свидетельствуют о высоком качестве компоненты нормализации именованных сущностей по сравнению с решениями других команд участников.

4 Заключение

В заключительном разделе представлены основные итоги работы.

1. В работах [1; 2] изучены классификационные модели применительно к задаче нормализации медицинских концептов. Предложены признаки семантической близости, которые показали свою эффективность в рамках открытых тестирований CLEF eHealth 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020 Task 3, SMM4H 2021 Task 1c. Также в работе показаны недостатки классификационного подхода и стандартных методов оценки качества моделей. В частности, отсутствие обучающих выборок, покрывающих все возможные медицинские концепты и значительные пересечения между проверочной и обучающей подвыборок существующих наборов данных. Предложен метод оценки моделей, устраняющий недостаток высокого пересечения.
2. Предложена модель DILBERT для связывания именованных сущностей, основанная на метрическом подходе. В работе [5] показана устойчивость данной модели к смене словаря и способность распознавать концепты, не присутствовавшие в обучающей выборке. Показана эффективность модели на корпусах CDR Disease, CDR Chemical, CT Intervention, CT Condition. Также показано, что модель DILBERT менее склонна к запоминанию словаря.
3. Исходя из особенностей классификационного и метрического подходов, предложена комбинированная модель связывания именованных сущностей. Показана эффективность данного метода в рамках соревнований SMM4H 2020 Task 3 и SMM4H 2021 Task 1c.

Литература

1. Medical concept normalization in social media posts with recurrent neural networks / Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, Valentin Malykh // *Journal of biomedical informatics*. — 2018. — Vol. 84. — Pp. 93–102.
2. *Miftahutdinov Zulfat, Tutubalina Elena*. Deep Neural Models for Medical Concept Normalization in User-Generated Texts // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. — 2019. — Pp. 393–399.
3. *Miftahutdinov Zulfat, Tutubalina Elena*. Deep learning for ICD coding: Looking for medical concepts in clinical documents in English and in French // International Conference of the Cross-Language Evaluation Forum for European Languages / Springer. — 2018. — Pp. 203–215.
4. *Miftahutdinov Z, Tutubalina E*. KFU at CLEF eHealth 2017 Task 1: ICD-10 coding of English death certificates with recurrent neural networks // CEUR Workshop Proceedings. — 2017.
5. Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer / Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina // *Proceedings of the 43rd European Conference on Information Retrieval*. — 2021.
6. Medical concept normalization in clinical trials with drug and disease representation learning / Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina // *Bioinformatics*. — 2021. — Vol. 37, no. 21. — Pp. 3856–3864.
7. *Miftahutdinov Zulfat, Sakhovskiy Andrey, Tutubalina Elena*. KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — 2020. — Pp. 51–56.
8. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019 / Davy Weissenbacher, Abeed Sarker, Arjun Magge et al. // Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task. — 2019. — Pp. 21–30.

9. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020 / Ari Klein, Ilseyar Alimova, Ivan Flores et al. // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — 2020. — Pp. 27–36.

10. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021 / Arjun Magge, Ari Klein, Antonio Miranda-Escalada et al. // Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task. — 2021. — Pp. 21–32.

11. *Canese Kathi, Weis Sarah.* PubMed: the bibliographic database // The NCBI Handbook [Internet]. 2nd edition. — National Center for Biotechnology Information (US), 2013.

12. *Aronson Alan R.* Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. // Proceedings of the AMIA Symposium / American Medical Informatics Association. — 2001. — P. 17.

13. *Leaman Robert, Islamaj Doğan Rezarta, Lu Zhiyong.* DNorm: disease name normalization with pairwise learning to rank // *Bioinformatics.* — 2013. — Vol. 29, no. 22. — Pp. 2909–2917.

14. *Leaman Robert, Lu Zhiyong.* TaggerOne: joint named entity recognition and normalization with semi-Markov Models // *Bioinformatics.* — 2016. — Vol. 32, no. 18. — Pp. 2839–2846.

15. *Limsopatham Nut, Collier Nigel.* Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2016. — Pp. 1014–1023.

16. Cadec: A corpus of adverse drug event annotations / Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, Chen Wang // *Journal of biomedical informatics.* — 2015. — Vol. 55. — Pp. 73–81.

17. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications / Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick et al. // *Data in brief.* — 2019. — Vol. 24. — P. 103838.

18. *Sarker Abeed, Gonzalez-Hernandez Graciela.* Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017 // *Training.* — 2017. — Vol. 1, no. 10,822. — P. 1239.

19. *Alvaro Nestor, Miyao Yusuke, Collier Nigel.* TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations // *JMIR public health and surveillance.* — 2017. — Vol. 3, no. 2. — P. e24.

20. *Doğan Rezarta Islamaj, Leaman Robert, Lu Zhiyong.* NCBI disease corpus: a resource for disease name recognition and concept normalization // *Journal of biomedical informatics.* — 2014. — Vol. 47. — Pp. 1–10.

21. Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — 2018. — Pp. 2227–2237.

22. *Miftahutdinov ZS, Tutubalina EV, Tropsha AE.* Identifying disease-related expressions in reviews using conditional random fields // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2017. — Pp. 155–166.

23. Hierarchical attention networks for document classification / Zichao Yang, Diyi Yang, Chris Dyer et al. // Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. — 2016. — Pp. 1480–1489.

24. Multi-task Character-Level Attentional Networks for Medical Concept Normalization / Jinghao Niu, Yehui Yang, Siheng Zhang et al. // *Neural Processing Letters.* — 2018. — Pp. 1–18.

25. Team UKNLP: Detecting ADRs, classifying medication intake messages, and normalizing ADR mentions on twitter / S. Han, T. Tran, A. Rios, R. Kavuluru // *CEUR Workshop Proceedings.* — 2017. — Vol. 1996. — Pp. 49–53.

26. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task / Abeed Sarker, Maksim Belousov, Jasper Friedrichs et al. // *Journal of the American Medical Informatics Association.* — 2018. — Vol. 25, no. 10. — Pp. 1274–1283.

27. LITL at CLEF eHealth2017: automatic classication of death reports / Lydia-Mai Ho-Dac, Cécile Fabre, Anouk Birski et al. // CLEF eHealth 2017. — 2017.

28. CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. / Aurélie Névéol, Aude Robert, Robert Anderson et al. // CLEF (Working Notes). — 2017.

29. *Cabot Chloé, Soualmia Lina Fatima, Darmoni Stéfan Jacques.* SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND. // CLEF (Working Notes). — 2017.

30. *Zweigenbaum Pierre, Lavergne Thomas.* Multiple Methods for Multi-class, Multi-label ICD-10 Coding of Multi-granularity, Multilingual Death Certificates. // CLEF (Working Notes). — 2017.

31. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation / Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre et al. // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2014. — Pp. 1724–1734.

32. *Bodenreider Olivier.* The unified medical language system (UMLS): integrating biomedical terminology. — Vol. 32. — Oxford University Press, 2004. — Pp. D267–D270.

33. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring / Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston // *CoRR abs/1905.01969. External Links: Link Cited by.* — 2019. — Vol. 2. — Pp. 2–2.

34. *Reimers Nils, Gurevych Iryna.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — Pp. 3973–3983.

35. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — Pp. 3111–3119.

36. *Tutubalina Elena, Kadurin Artur, Miftahutdinov Zulfat.* Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based

Models // Proceedings of the 28th International Conference on Computational Linguistics. — 2020. — Pp. 6710–6716.

37. Biomedical Entity Representations with Synonym Marginalization / Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, Jaewoo Kang // ACL. — 2020.