DOI: 10.15514/ISPRAS-2021-33(6)-15

Межъязыковой перенос знаний при извлечении информации о лекарствах из пользовательских текстов

[1, 2] А.С. Саховский, ORCID: 0000-0003-2762-2910 <andrey.sakhovskiy@gmail.com>
[2, 3, 4] Е.В. Тутубалина, ORCID: 0000-0001-7936-0284 <tutubalinaev@gmail.com>

[1] Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1.
[2] Казанский федеральный университет,
420008, Россия, Казань, ул. Кремлевская, д.18.
[3] Национальный исследовательский университет "Высшая школа экономики",
101000, Россия, Москва, ул. Мясницкая, д.20
[4] Sber AI
121170, Россия, Москва, Кутузовский пр-кт, д. 32

Аннотация. Задача извлечения именованных сущностей, соответствующих лекарствам, заболеваниям и лекарственным реакциям, из текстов различных предметных областей и языков является основополагающим компонентом многих приложений, основанных на извлечении информации из текстов. В данной работе производится оценка эффективности многоязыковых моделей, основанных на архитектуре BERT, для решения задач распознавания именованных сущностей медицинской направленности и многоклассовой классификации предложений. В ходе экспериментов было исследовано влияние переноса знаний между двумя англоязычными корпусами и одним русскоязычным корпусом размеченных отзывов о лекарственных препаратах. Рассмотренные корпуса содержат разметку на уровне предложений, обозначающую присутствие или отсутствие в них медицинских сущностей некоторого типа. Предложения, принадлежащие некоторому классу, содержат дополнительную разметку на уровне сущностей, позволяющую установить принадлежность отдельных выражений к сущностям некоторого типа, таким, как название, показание к применению или эффект лекарства. Результаты экспериментов показали, что для русского языка наибольшая эффективность переноса знаний при предобучении моделей BERT на коллекции, состоящей из 5 миллионов неразмеченных русскоязычных и англоязычных пользовательских отзывах, наблюдается при распознавании побочных эффектов лекарств. Для задачи распознавания именованных сущностей наилучшее значение макро F-меры, равное 74,85%, показала модель RuDR-BERT, предобученная на русскоязычных текстах медицинской предметной области. Для задачи классификации наилучшее значение макро F-меры, равное 70%, показала модель EnRuDR-BERT, предобученная на русскоязычных и англоязычных текстах медицинской направленности. Превосходство данной модели над моделью BERT, предобученной на текстах общей предметной области, составляет 8,64% макро F-меры.

Ключевые слова: обработка естественного языка; классификация текстов; извлечение информации; распознавание именованных сущностей; BERT

Для цитирования: Саховский А.С., Тутубалина Е.В. Межъязыковой перенос знаний при извлечении информации о лекарствах из пользовательских текстов. Труды ИСП РАН, том 33, вып. 6, 2021 г., стр. 217-228. DOI: 10.15514/ISPRAS–2021–33(6)–15

Благодарности: Данная работа выполнена при поддержке гранта Президента РФ МК-3193.2021.1.6

Cross-lingual transfer learning in drug-related information extraction from user-generated texts

[1, 2] A.S. Sakhovskiy ORCID: 0000-0003-2762-2910 <andrey.sakhovskiy@gmail.com>
[2, 3, 4] E.V. Tutubalina ORCID: 0000-0001-7936-0284 <tutubalinaev@gmail.com>

[1] Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.
[2] Kazan Federal University,
18 Kremlyovskaya street, Kazan, 420008, Russia.
[3] National research university Higher school of economics,
20 Myasnitskaya street, Moscow, 101000, Russia.
[4] Sber AI,
32 Kutuzovskiy prospect, Moscow, 121170, Russia

Abstract. Aggregating knowledge about drug, disease, and drug reaction entities across a broader range of domains and languages is critical for information extraction (IE) applications. In this work, we present a fine-grained evaluation intended to understand the efficiency of multilingual BERT-based models for biomedical named entity recognition (NER) and multi-label sentence classification tasks. We investigate the role of transfer learning (TL) strategies between two English corpora and a novel annotated corpus of Russian reviews about drug therapy. Labels for sentences include health-related issues or their absence. The sentences with one are additionally labelled at the expression level to identify fine-grained subtypes such as drug names, drug indications, and drug reactions. Evaluation results demonstrate that BERT trained on Russian and English raw reviews (5M in total) shows the best transfer capabilities on evaluation of adverse drug reactions on Russian data. The macro F1 score of 74.85% in the NER task was achieved by our RuDR-BERT model. For the classification task, our EnRuDR-BERT model achieves the macro F1 score of 70%, gaining 8.64% over the score of a general domain BERT model.

Keywords: natural language processing; text classification; information extraction; named entity recognition; BERT.

For citation: Sakhovskiy A.S., Tutubalina E.V. Cross-lingual transfer learning in drug-related information extraction from user-generated texts. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 6, 2021, pp. 217-228 (in Russian). DOI: 10.15514/ISPRAS–2021–33(6)–15

Acknowledgements. The work has been supported by a grant from the President of the Russian Federation for young scientists-candidates of science (MK-3193.2021.1.6).

1. Введение

Значительная часть существующих работ в области обработки текстов биомедицинской тематики посвящена обработке англоязычных текстов. В частности, обработке англоязычных научных текстов, например, научных аннотаций. В работе [1] произведен обзор данной предметной области для английского языка. Однако задача обработки пользовательских текстов биомедицинской тематики для языков, отличных от английского, в настоящий момент слабо изучена. Значительные продвижения в области разработки многоязыковых нейросетевых моделей обработки текстов, в частности, предобученных языковых моделей, основанных на архитектуре Transformer [2], позволяют получать информативные векторные представления слов, зависящие от окружающего словесного контекста [3, 4, 5, 6]. Данные продвижения позволяют выдвинуть предположение о возможности улучшения качества и разработки более эффективных моделей для широкого круга задач обработки естественного языка, включая классификацию текстов и извлечение именованных сущностей в биомедицинской предметной области.

Одним из значимых направлений современных исследований в области обработки естественного языка является разработка методов так называемого переноса знаний, суть которого состоит в использовании информации, полученной некоторой моделью при

решении одной задачи, для решения некоторой другой. Одним из проявлений переноса знаний может являться использование предобученных языковых моделей. Идея предобучения языковых моделей состоит в обучении некоторой модели на огромном неразмеченном наборе данных с целью извлечения из данного набора данных информации, которая в дальнейшем может быть использована для решения другой задачи. Результатом предобучения является некоторая предобученная модель, позволяющая получать информативные векторные представления текста и его отдельных слов.

Ввиду неравномерного развития ресурсной базы для различных языков, особую значимость приобретает задача межъязыкового переноса знаний, состоящая в использовании знаний и ресурсов, разработанных для одного языка, при решении задач на некотором другом языке.

В отличие от предшествующих исследований, рассматривавших перенос знаний в рамках текстов одного языка, ключевым аспектом данной работы является изучение эффективности межъязыкового переноса знаний в области биомедицинских текстов. В рамках данной работы была проведена оценка эффективности (i) различных предобученных моделей архитектуры BERT при переносе знаний внутри русского языка, а также (ii) методов межъязыкового переноса знаний из английского языка в русский применительно к текстам медицинской направленности. В качестве задач были рассмотрены (i) задача классификации предложений и (ii) извлечение именованных сущностей применительно к пользовательским текстам медицинской тематики. Целью данного исследования является оценка возможности использования английского языка, для которого существует развитая ресурсная база, для улучшения качества извлечения информации о лекарствах и заболеваниях из текстов на русском языке. Для проведения экспериментов по оценке данной возможности в данной работе используется русскоязычный корпус Russian Drug Reaction Corpus (RuDReC) [7], предназначенный для исследований в области извлечения из текстов информации о лекарствах и их побочных эффектах, а также о заболеваниях. Данная работа продолжает исследование эффективности переноса знаний для русского языка, начатое в работе [7].

2. Обзор предметной области

В последние годы значительная часть работ, посвященных задачам извлечения информации из текстов о лекарствах, основана на использовании англоязычных текстов социальных сетей, пользовательских отзывов и медицинских записей [8, 9, 10, 11]. Однако существует небольшое число работ, посвященных другим языкам. Так, в работе [12] были рассмотрены описания лекарств на испанском языке, а в работах [13, 14] были использованы сертификаты о смерти на французском языке. Что касается русского языка, в работе [15] были рассмотрены клинические записи на русском языке, а в работах [16, 17] – русскоязычные записи пользователей социальной сети Twitter.

В отличие от английского языка, число доступных размеченных корпусов, посвященных извлечению информации о лекарствах и заболеваниях, для русского языка относительно мало. Среди существующих корпусов можно выделить следующие. В работе [18] был представлен корпус русскоязычных отзывов о лекарственных препаратах, содержащий разметку на уровне предложений. Корпус предоставляет разметку по 4 классам в зависимости об информации, содержащейся в предложении, включая побочные эффекты, показания к применению и информацию об эффективности лекарственного препарата. Отдельно необходимо выделить русскоязычные корпуса соревнований Social Media Mining for Health Applications (SMM4H) 2020 и 2021 годов [19, 20]. В рамках данных соревнований были представлены корпуса русскоязычных твитов, размеченных на предмет содержания в них упоминаний побочных эффектов.

3. Данные

В данной работе было проведено исследование эффективности предобученных моделей архитектуры BERT для решения задач классификации текстов и извлечения именованных сущностей медицинской тематики на русском языке. Для обучения и оценки качества моделей был использован корпус RuDReC. The Russian Drug Reaction Corpus [7] (RuDReC) – корпус русскоязычных пользовательских отзывов о лекарственных препаратах. Корпус доступен для исследовательских целей по ссылке: https://github.com/cimm-kzn/RuDReC. Данный корпус состоит из двух частей: (i) неразмеченной части, состоящей из 1,4 миллионов текстов и (ii) размеченной части, включающей 500 текстов. Отзывы размеченной части корпуса содержат разметку как на уровне предложений, так и на уровне сущности. Каждое предложение может принадлежать одному или нескольким классам предложений из следующего списка:

- предложение, в котором сообщается о положительном эффекте приема лекарственного препарата (класс DE);
- предложение, в котором сообщается о том, что в результате приема лекарственного препарата состояние пациента осталось неизменным или ухудшилось (класс DIE);
- предложение, в котором содержится упоминание симптомов и показаний к применению, побудивших пользователя к приему лекарства (класс DI);
- предложение, в котором содержится упоминание нежелательных побочных эффектов, возникших в результате приема лекарства (класс ADR);
- предложение, содержащее упоминание некоторого события или эффекта, связанного с болезнью или лекарственным препаратов (класс Finding). Предложения данного типа могут содержать историю болезни пациента, название лекарства, сообщать об отсутствии ожидаемого побочного эффекта.

Статистика по числу размеченных предложений каждого класса приведена в табл. 1. Как можно заметить, наибольшее число предложений представлено для класса симптомов (DI), а наименьшим числом примеров обладает класс Finding.

Табл. 1. Статистика по числу размеченных предложений различных классов корпуса RuDReC
Table 1. RuDReC corpus statistics on number of sentences of different classes

Общее число предложений	DE	DIE	ADR	DI	Finding
4855	424	278	379	949	172

Помимо разметки на уровне предложений тексты корпуса RuDReC содержат разметку на уровне сущностей, что позволяет использовать данный корпус для обучения и оценки моделей извлечения именованных сущностей медицинской тематики. Размеченные сущности принадлежат одному из следующих 6 возможных типов: (i) ADR, (ii) DI, (iii) Finding, (iv) Drugclass (класс лекарственного препарата. Например, противовирусный или противовоспалительный препарат и др.), (v) Drugform (лекарственная форма препарата. Например, таблетки, микстура, мазь и др.), (vi) Drugname (название лекарства или его действующего вещества).

Помимо русскоязычного корпуса RuDReC, в данной работе были использованы англоязычные корпуса PsyTAR [9] и CADEC [10]. Psychiatric Treatment Adverse Reactions (PsyTAR) – корпус, состоящий из 887 размеченных пользовательских отзывов о психиатрических препаратах. По структуре и содержанию разметки данный корпус имеет высокое сходство с размеченной частью корпуса RuDReC. Так, данный корпус содержит разметку как на уровне предложений, так и на уровне сущностей. Согласно разметке на уровне предложений, каждое предложение может быть отнесено к одному или нескольким из 7 классов. Набор возможных классов содержит классы, аналогичные классам DE, DIE,

ADR, DI, Finding. Основное отличие в наборе классов корпуса PsyTAR от набора корпуса RuDReC состоит в наличии двух дополнительных классов предложений: (i) сообщающих о том, что прекращение использования лекарства привело к возникновению у пользователя синдрома отмены (класс WD); (ii) сообщающих о некотором испытанном пользователем симптоме, не являющемся ни причиной применения, ни побочным эффектом приема лекарства (класс SSI). В рамках данной работы при обучении моделей на корпусе PsyTAR была использована разметка только по тем классам предложений, для которых существует аналогичный класс в корпусе RuDReC, т.е. классы WD и SSI не рассматривались. Разметка на уровне сущностей содержит включает сущности 4 типов: ADR, WD, DI, SSI.

В табл. 2 представлена статистика по числу размеченных предложений корпуса PsyTAR. Приведенная статистика позволяет сделать следующие основные наблюдения. Во-первых, наиболее частотным классом предложений является класс предложений, содержащих упоминание побочных эффектов. Число предложений данного класса в корпусе PsyTAR более, чем в 5 раз выше, чем в корпусе RuDReC (2168 и 379 для корпусов PsyTAR и RuDReC соответственно). Во-вторых, число предложений класса Finding более, чем в 10 больше в корпусе PsyTAR, чем в корпусе RuDReC (2107 и 172 соответственно).

Табл. 2. Статистика по числу размеченных предложений различных классов корпуса PsyTAR
Table 2. PsyTAR corpus statistics on number of sentences of different classes

Общее число предложений	DE	DIE	ADR	DI	Finding
6004	1087	337	2168	517	2107

Корпус CSIRO Adverse Drug Event Corpus (CADEC) состоит из 1253 пользовательских отзывов о 12 лекарственных препаратах. Данный корпус предоставляет разметку сущностей 5 типов: (i) название лекарства, (ii) ADR, (iii) заболевание и (iv) его симптомы, побудившие пользователя к приему лекарства, (v) Finding. В данной работе мы рассматриваем сущности классов (iii) и (iv) как сущности одного общего класса DI.

4. Модели

В данной работе были использованы следующие предобученные модели архитектуры Bidirectional Encoder Representations from Transformers (BERT) [3]:

1) Multi-BERT[1] – многоязыковая модель BERT, предобученная на текстах Википедии на 104 языках.
2) RuBERT [21] – многоязыковая модель BERT, дополнительно предобученная на русскоязычных текстах Википедии и новостных текстах. Для инициализации данной модели была использована модель Multi-BERT с измененным словарем
3) RuDR-BERT[2] [7] – многоязыковая модель BERT, дополнительно предобученная на неразмеченной части корпуса RuDReC [7], содержащей 1,4 миллионов отзывов о лекарственных препаратах. В качестве инициализации для данной модели была использована модель Multi-BERT;
4) EnRuDR-BERT[3] – многоязыковая модель BERT, предобученная на неразмеченной части корпуса RuDReC и англоязычном корпусе пользовательских текстов о лекарствах [22]. Объем англоязычного корпуса составляет около 2,6 миллионов текстов.

Каждая из использованных моделей содержит 12 слоев, 12 голов интерактивного внимания и имеет размерность векторных представлений, равную 768. В рамках данной работы для

[1] https://github.com/google-research/bert
[2] https://huggingface.co/cimm-kzn/rudr-bert
[3] https://huggingface.co/cimm-kzn/enrudr-bert

проведения экспериментов с предобученными моделями архитектуры BERT была использована программная реализация оригинальной модели BERT, доступная по адресу: https://github.com/google-research/bert.

5. Эксперименты

5.1. Классификация предложений

В данной работе для задачи классификации предложений было исследовано влияние переноса знаний на качество решения задачи на русском языке. В ходе экспериментов было исследовано влияние следующих способов переноса знаний: (i) перенос знаний в рамках одного языка путем предобучения языковых моделей BERT на текстах целевого языка и целевой предметной области; (ii) перенос знаний путем последовательного обучения модели на текстах вспомогательного и целевого языка на схожей задаче; (iii) перенос знаний путем обучения модели на вспомогательном языке и оценке качества классификации на целевом языке без обучения на текстах целевого языка (zero-shot перенос). Подход (iii) позволяет решать задачу на языках, для которых не существует размеченной тренировочной выборки за счет использования размеченной выборки на другом языке. В данной работе в качестве вспомогательного языка рассматривается английский, а в качестве целевого – русский язык.

В экспериментах, посвященных оценке влияния предобучения на русском языке, было произведено сравнение моделей Multi-BERT, RuBERT, RuDR-BERT, EnRuDR-BERT. Для обучения моделей был использован русскоязычный корпус RuDReC. При оценке влияния последовательного обучения сначала на английских, а затем на русских данных, были использованы англоязычный корпус PsyTAR и корпус RuDReC соответственно. Данные корпуса обладают схожими структурой разметки и набором возможных классов предложений, что позволяет переиспользовать одну и ту же модель для обучения на обоих корпусах без внесения изменений в ее архитектуру. В рамках данной серии экспериментов были рассмотрены русскоязычная модель RuDR-BERT и англо-русская модель EnRuDR-BERT. Эксперименты по оценке качества классификации при обучении на вспомогательном (английском) языке без обучения на целевом (русском) состояли в обучении моделей только на англоязычных текстах корпуса PsyTAR и оценке на корпусе RuDReC без дополнительного обучения на русскоязычных текстах. В рамках данных экспериментов также были рассмотрены модели RuDR-BERT и EnRuDR-BERT.

Для оценки эффективности каждого из рассмотренных подходов была использована процедура скользящего контроля с числом разбиений, равным 5. Каждый классификатор состоит из некоторой модели BERT и полносвязной нейронной сети. Полносвязная нейронная сеть принимает на вход векторное представление специального токена начала предложения. Обучение каждой модели происходило в течение 10 тренировочных эпох с кросс-энтропийной функцией потерь. Для сравнения качества различных обученных моделей классификации были использованы размеченные предложения русскоязычного корпуса RuDReC. В качестве метрики оценки качества моделей использована F-мера. Результаты соответствующих экспериментов представлены в табл. 3.

На основе полученных результатов можно сделать следующие основные наблюдения. Во-первых, многоязыковая модель EnRuDR-BERT значительно превзошла RuDR-BERT по макро F-мере (+7,3%) при обучении только на англоязычных данных. Наибольший прирост качества наблюдается для предложений, содержащих упоминания побочных эффектов (+28,6%) и предложений, сообщающих о положительном эффекте лекарства (+9,8%). Для предложений класса Finding превосходство модели EnRuDR-BERT незначительно, а для предложений, сообщающих о неэффективности лекарства и для предложений, содержащих упоминания симптомов, данная модель и вовсе уступила модели RuDR-BERT на 2,24% и 0,73% F-меры соответственно.

Табл. 3. Оценки F-меры предобученных моделей BERT на задаче классификации предложений корпуса RuDReC
Table 3. Performance of pretrained BERT models on the classification of RuDReC corpus sentences in terms of F1-score

Модель	DE	DIE	ADR	DI	Finding	Макро F-мера
Модели, обученные только на англоязычном корпусе PsyTAR (zero-shot перенос)						
RuDR-BERT, PsyTAR	41,69	59,91	36,29	18,05	2,22	31,63
EnRuDR-BERT, PsyTAR	51,51	57,67	64,93	17,32	3,15	38,92
Модели, обученные только на корпусе RuDReC						
RuBERT	67,7	62,27	66,65	81,63	28,51	61,35
Multi-BERT	63,61	60,19	63,45	79,58	24,32	58,23
RuDR-BERT	76,61	72,06	74,15	85,06	36,24	68,82
EnRuDR-BERT	78,01	74,47	**75,54**	85,37	35,47	69,77
Модели, последовательно обученные на английских и русских данных						
RuDR-BERT, CADEC+RuDReC	77,72	**75,78**	74,14	85,69	33,86	69,44
RuDR-BERT, PsyTAR+RuDReC	77,87	73,5	74,32	85,5	30,7	68,38
EnRuDR-BERT, PsyTAR+RuDReC	77,68	71,99	75,43	85,62	**39,3**	**70,0**
EnRuDR-BERT, CADEC+RuDReC	**78,58**	72,19	75,51	**86,31**	36,71	69,86

Во-вторых, при обучении и оценке на русскоязычных данных наивысшие значения F-меры демонстрируют модели EnRuDR-BERT и RuDR-BERT, предобученные на русскоязычных текстах целевой медицинской предметной области. В частности, модель EnRuDR-BERT, предобученная на англо- и русскоязычных текстах, превзошла русскоязычную модель RuBERT на 8,4% макро F-меры, а наибольший прирост F-меры наблюдается для классов DE (+10,3%), DIE (+12,2%), ADR (+8,9%). Наихудшие значения F-меры для всех классов показала многоязыковая модель Multi-BERT, которая, в отличие от остальных рассмотренных моделей, не проходила дополнительное предобучение на русскоязычных текстах. Данное наблюдение свидетельствует об эффективности предобучения на текстах целевого (русского) языка. Однако на эффективность предобучения существенно влияет предметная область текстов, на которых происходит предобучение. Так, несмотря на превосходство модели RuBERT, предобученной на русскоязычных новостных текстах, над моделью Multi-BERT, данная модель уступает моделям RuDR-BERT и EnRuDR-BERT, прошедшим предобучение на текстах медицинской тематики, то есть на текстах той же предметной области, что и тексты целевой задачи классификации.

В-третьих, последовательное обучение на англоязычных и русскоязычных данных не привело к существенному улучшению качества классификации с точки зрения макро F-меры, однако последовательное обучение модели EnRuDR-BERT на корпусах PsyTAR и RuDReC

привело к увеличению F-меры класса Finding на 3,8% по сравнению с обучением только на корпусе RuDReC. Наконец, для всех проведенных экспериментов наихудшие оценки качества наблюдаются для класса Finding, что может объясняться малым количеством тренировочных примеров данного класса.

5.2. Распознавание именованных сущностей

По аналогии с задачей классификации предложений, в данной работе было исследовано влияние предобучения на качество решения решения задачи распознавания именованных сущностей медицинской тематики. В ходе экспериментов было произведено сравнение следующих моделей: (i) многоязыковой модели Multi-BERT; многоязыковых моделей (ii) RuBERT и (iii) RuDR-BERT, прошедших предобучение на русскоязычных текстах; (iv) многоязыковой модели EnRuDR-BERT. Кроме того, была проведена оценка эффективности переноса знаний из английского языка путем последовательного обучения сначала на вспомогательном англоязычном корпусе с последующим дообучением на корпусе RuDReC.

Для оценки качества моделей была использована процедура скользящего контроля с 5 разбиениями. Архитектура классификатора состоит из предобученной модели BERT с дополнительным слоем softmax. Обучение каждой модели происходило в течение 40 тренировочных эпох. Результаты соответствующих экспериментов представлены в таблице 4.

Полученные результаты позволяют сделать следующие основные наблюдения. Во-первых, как и в случае задачи классификации предложений модели RuDR-BERT и EnRuDR-BERT, предобученные на текстах медицинской предметной области, показали наилучшие значения макро F-меры. Данное наблюдение позволяет сделать вывод об эффективности предобучения на текстах предметной области и языка целевой задачи. Во-вторых, аналогично задаче классификации, качество распознавания сущностей типа Finding значительно ниже, чем сущностей побочных эффектов и симптомов. В-третьих, качество распознавания сущностей, связанных с лекарствами – их названиями, классами и лекарственными формами – значительно превосходит качество распознавания сущностей, связанных с заболеваниями – ADR, DI и Finding. Данное наблюдение может объясняться меньшей длиной сущностей, связанных с лекарствами. Так, средняя длина сущностей Drugclass, Drugform и Drugname в корпусе RuDReC составляет 1,06 слов, а сущностей ADR, DI и Finding – 1,77. Таким образом, распознавание сущностей последних трех типов представляет собой более трудную задачу, поскольку такие сущности зачастую являются словосочетаниями. Наконец, результаты экспериментов показали, что использование англоязычных данных в процессе обучения модели приводит не к повышению, а понижению качества распознавания с точки зрения макро F-меры. Тем не менее, модель RuDR-BERT, последовательно обученная на PsyTAR и RuDReC, показала прирост F-меры при распознавании сущностей побочных эффектов (+2,1%).

Табл. 4. Оценки F-меры предобученных моделей BERT на задаче извлечения именованных сущностей корпуса RuDReC
Table 4. Performance of pretrained BERT models on the RuDReC corpus named entity recognition task

Модель	ADR	DI	Finding	Drugclass	Drugform	Drugname	Макро F-мера
RuBERT	54.51	69.43	27.87	92.78	95.72	92.11	72.07
Multi-BERT	54.65	67.63	25.75	92.36	94.89	91.05	71.06
RuDR-BERT	60.36	72.33	**33.31**	**94.12**	95.89	93.08	**74.85**

EnRuDR-BERT	61.11	**72.84**	27.67	94.08	**96.20**	92.18	74.01
RuDR-BERT, CADEC+RuDReC	57.74	71.43	28.94	93.08	95.38	93.01	73.26
RuDR-BERT, PsyTAR+RuDReC	**62.48**	71.78	28.60	92.67	95.28	**93.17**	74.00

6. Заключение

В результате данной работы были получены следующие основные результаты. Во-первых, было исследовано влияние предметной области текстовой коллекции, использованной для предобучения модели BERT, на качество решения двух задач на русском языке: многоклассовой классификации предложений и распознавания именованных сущностей названий лекарств, показаний к применению и побочных эффектов. В ходе экспериментов была проведена оценка эффективности двух моделей архитектуры BERT: (i) RuDR-BERT, предобученной на неразмеченной части русскоязычного корпуса отзывов о лекарственных препаратах RuDReC и (ii) EnRuDR-BERT, преобученной на объединении RuDReC с англоязычным неразмеченным корпусом пользовательских отзывов медицинской тематики. В ходе экспериментов было показано, что данные модели превосходят как многоязыковую модель Multi-BERT, так и русскоязычную модель RuBERT, обученную на текстах общей предметной области. Таким образом, результаты проведенных экспериментов позволяют сделать вывод об эффективности переноса знаний путем предобучения языковых моделей BERT на неразмеченных данных медицинской предметной области.

Во-вторых, была исследована возможность использования англоязычных данных для улучшения качества решения рассмотренных задач на русском языке. Результаты показали, что последовательное обучение сначала на англоязычных данных, а затем на русскоязычных не приводит к существенному общему падению качества с точки зрения макро F-меры, при этом для отдельных типов сущностей наблюдается даже улучшение качества. Так, в задаче извлечения именованных сущностей, соответствующих побочным эффектам лекарств, последовательное обучение на англоязычном корпусе PsyTAR и корпусе RuDReC привело к увеличению F-меры на 2,1% по сравнению с обучением только на корпусе RuDReC.

В-третьих, было проведено сравнение моделей RuDR-BERT и EnRuDR-BERT при обучении только на англоязычных данных с последующей оценкой качества на корпусе RuDReC. В рамках данного эксперимента двуязыковая модель EnRuDR-BERT превзошла русскоязычную модель RuDR-BERT на 7,3% макро F-меры в задаче классификации предложений корпуса RuDReC, в то время как при обучении на RuDReC превосходство EnRuDR-BERT составляет лишь 1% макро F-меры. Данное наблюдение позволяет сделать вывод об эффективности использования размеченных данных одного языка при решении схожей задачи для другого языка, вовсе не имеющего размеченного корпуса.

Список литературы / References

[1]. Huang C.C., Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics, vol. 17, no. 1, 2016, pp. 132-144.

[2]. Vaswani A., Shazeer N. et al. Attention is all you need. In Proc. of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000-6010.

[3]. Devlin J., Chang M. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (Long and Short Papers), 2019, pp. 4171-4186.

[4]. Conneau A., Lample G. Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 7059-7069.

[5]. Lample G., Conneau A. et al. Unsupervised Machine Translation Using Monolingual Corpora Only. In Proc. of the International Conference on Learning Representations, 2018, 14 p.

[6]. Artetxe M., Schwenk H. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3197-3203.

[7]. Tutubalina E., Alimova I. et al. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. Bioinformatics, vol. 37, issue 2, 2021, pp. 243-249.

[8]. Alvaro N., Miyao Y., Collier N. TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. JMIR public health and surveillance, vol. 3, issue 2, 2017, article id. e6396.

[9]. Zolnoori M. et al. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for SSRI and SNRI medications. Journal of biomedical informatics, vol. 90, 2019, article no. 103091.

[10]. Karimi S., Metke-Jimenez A. et al. Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics, vol. 55, 2015, pp. 73-81.

[11]. Sarker A., Belousov M. et al. 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. Journal of the American Medical Informatics Association, vol. 25, issue 10, 2018, pp. 1274-1283.

[12]. Moreno I., Boldrini E. et al. Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics. Journal of biomedical informatics, vol. 72, 2017, pp. 8-22.

[13]. Névéol A., Anderson R.N. et al. CLEF eHealth 2017 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in English and French. CLEF 2017 Working Notes. CEUR Workshop Proceedings, vol. 1866, 2017, 17 p.

[14]. Névéol A. et al. CLEF eHealth 2018 Multilingual Information Extraction Task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. CLEF 2018 Working Notes. CEUR Workshop Proceedings, vol. 2125, 2018, 18 p.

[15]. Shelmanov A.O., Smirnov I.V., Vishneva E.A. Information extraction from clinical texts in Russian. In Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, issue 14, 2015, pp. 560-572.

[16]. Miftahutdinov Z., Sakhovskiy A., Tutubalina E. Kfu nlp team at smm4h 2020 tasks: Cross-lingual transfer learning with pretrained language models for drug reactions. In Proc. of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020, pp. 51-56.

[17]. Gusev A., Kuznetsova A. et al. Bert implementation for detecting adverse drug effects mentions in russian In Proc. of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020, pp. 46-50.

[18]. Alimova I., Tutubalina E. et al. A Machine learning approach to classification of drug reviews in Russian. In Proc. of the Ivannikov ISPRAS Open Conference, 2017, pp. 64-69.

[19]. Klein A., Alimova I. et al. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In Proc. of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020, pp. 27-36.

[20]. Magge A., Klein A. et al. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at NAACL 2021. In Proc. of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021, pp. 21-32.

[21]. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. arXiv preprint arXiv:1905.07213, 2019.

[22]. Тутубалина Е. В., Мифтахутдинов З. Ш. и др. Идентификация лекарственных средств со схожим терапевтическим действием на основе семантического анализа текстов. Известия академии наук. Серия химическая, no. 11, 2017 г., стр. 2180-2189 / Tutubalina E.V., Miftahutdinov Z. Sh. et al. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. Russian Chemical Bulletin, vol. 66. issue 11, 2017, pp. 2180-2189.

Информация об авторах / Information about authors

Елена Викторовна ТУТУБАЛИНА – кандидат физико-математических наук, научный сотрудник НУЛ "Моделей и методов вычислительной прагматики" факультета компьютерных наук Высшей школы экономики; старший научный сотрудник НИЛ "Хемоинформатика и молекулярное моделирование" Казанского федерального университета; исполнительный директор по исследованию данных в Sber AI. Сфера научных интересов: обработка текстов на естественном языке, извлечение информации из текстов, распознавание именованных сущностей, классификация текстов.

Elena Viktorovna TUTUBALINA – Candidate of Physical and Mathematical Sciences, Researcher at the "Models and Methods of Computational Pragmatics" Research Laboratory of the Faculty of Computer Science, Higher School Economics; Senior Researcher at the "Chemoinformatics and Molecular Modeling" Research Laboratory, Kazan Federal University; Executive Director of Data Science at Sber AI. Her research interests include natural language processing, information extraction, named entity recognition, text classification.

Андрей Сергеевич САХОВСКИЙ – лаборант НИЛ "Хемоинформатика и молекулярное моделирование" Казанского федерального университета; студент 1 курса кафедры математических методов прогнозирования факультета вычислительной математики и кибернетики Московского государственного университета. Сфера научных интересов: обработка текстов на естественном языке, классификация текстов, анализ текстов социальных сетей.

Andrey Sergeyevich SAKHOVSKIY – Laboratory Assistant at the "Chemoinformatics and Molecular Modeling" Research Laboratory of the Kazan Federal University; 1st-year graduate student of the Department of Mathematical Forecasting Methods, Faculty of Computational Mathematics and Cybernetics, Moscow State University. His research interests include natural language processing, text classification, social media text analysis.