УДК 615.074+542.8:539.19+004.912

Идентификация лекарственных средств со схожим терапевтическим действием на основе семантического анализа текстов*

Е. В. Тутубалина,[а] З. Ш. Мифтахутдинов,[а] Р. И. Нугманов,[а] Т. И. Маджидов,[а] С. И. Николенко,[а,б] И. С. Алимова,[а] А. Э. Тропша[а,в]*

[а]*Казанский (Приволжский) федеральный университет, Российская Федерация, 420008 Казань, ул. Кремлевская, 18. E-mail: elvtutubalina@kpfu.ru*
[б]*Санкт-Петербургское отделение Математического института им. В. А. Стеклова Российской академии наук, Российская Федерация, 191011 Санкт-Петербург, Реки Фонтанки наб., 27*
[в]*Университет Северной Каролины в Чапел-Хилл, США, Северная Королина, 27514 Чапел-Хилл, Кантри Клаб Роад, Джексон-Холл, 153А***

Описан подход к идентификации лекарственных средств со схожей терапевтической активностью на основе семантического анализа коллекции текстов. С помощью методов обработки естественного языка проанализировано >2.5 млн текстов отзывов о приеме лекарств на английском языке, опубликованных на форумах пациентов и в дискуссионных группах. Для получения векторного представления слов на основе входных данных построена модель Continuous Bag-of-Words, являющаяся одним из инструментов анализа семантики естественного языка word2vec. Это позволило каждому названию лекарственного препарата поставить в соответствие числовой вектор. После этого составлен список пар лекарственных средств, имеющих высокие значения близости полученных векторов. Анализ этого списка подтвердил, что наиболее подобные векторы соответствуют либо лекарствам с одинаковым действующим веществом, либо соединениям с близким терапевтическим эффектом и относящимся к одной и той же терапевтической группе. Обнаружено, что при этом часто химическое подобие веществ невелико. Предложенная технология использована для визуализации химического пространства лекарств и поиска веществ с потенциально подобным биологическим эффектом среди соединений различных терапевтических групп.

Ключевые слова: подобие химических соединений, перепрофилирование лекарств, обработка естественного языка, косинусная мера близости, векторные представления слов, word2vec, семантический анализ текстов, анализ отзывов пользователей, химическое пространство, хемоинформатика.

В настоящее время поиск гипотез о возможности перепрофилирования существующих лекарств с помощью анализа текстовой информации[1—3] представляет собой одно из перспективных направлений исследований академических научных групп и фармацевтических компаний. Классическим источником информации служат медицинские и биомедицинские научные тексты. Дополнительным современным источником текстовой информации являются мнения пациентов в интернет-пространстве (в социальных сетях, в дискуссионных группах, на форумах), поскольку пациенты на различных форумах и в социальных сетях обсуждают применение лекарственных препаратов для лечения заболеваний, в том числе использование лекарственных средств «off-label» (не по инструкции).

Число публикаций, посвященных анализу текстов интернет-пространства для решения задач медико-фармакологического анализа, стремительно растет с 2010 г. В работах[4—7] упоминается автоматизированная система «фармаконадзор» (англ. pharmacovigilance), одна из задач которой — мониторинг социальных сетей для идентификации потенциально неблагоприятных эффектов и взаимодействий лекарственных средств. Опубликован обзор[8] методов анализа естественного языка при идентификации реакций на лекарственный препарат, при этом в качестве источников текстовых данных рассматривались научные статьи, обсуждения в социальных сетях, электронные карты пациентов (англ. electronic health records). Выполнен анализ[9] различных методов обработки текстов для биомедицинских целей. Изучена[10] автоматическая обработка текстов научных статей и отзывов пациентов с целью извлечения новых взаимосвязей между лекарствами и болезнями. Отметим, что при использовании метода[11], основанного на словарях, для

* По материалам XX Менделеевского съезда по общей и прикладной химии (26—30 сентября 2016 г., Екатеринбург).
** University of North Carolina at Chapel Hill, Jackson Hall 153A, Country Club Road, Chapel Hill, NC 27514, North Carolina, USA.

идентификации из текстов пациентов с форума WebMD новых показаний к применению лекарств с целью перепрофилирования качество его работы не оценивалось.

Рассмотрена[12–16] задача извлечения новых побочных реакций лекарств из отзывов в социальных сетях с целью их дальнейшего перепрофилирования. Наиболее широко применяемый для решения данной задачи метод — это подход, основанный на словарях[17–24]. Словари состоят из списков побочных реакций, извлеченных из инструкций по применению лекарств, и записей о клинических испытаниях. Существуют также основанные на правилах методы[25,26], позволяющие выделять наиболее распространенные конструкции предложений, которые могут свидетельствовать об описании побочных реакций. В большинстве работ описаны исследования с помощью методов машинного обучения. Например, используются метод опорных векторов (SVM)[27–33], метод условных случайных полей (CRF)[34,35] и метод случайных лесов (Random Forest)[36]. Кроме того, побочные реакции выявляли[37] с помощью нейронных сетей.

Изучено[38] применение распределенных представлений слов для повышения качества извлечения информации из клинических заметок. Для этой цели в существующую модель машинного обучения интегрируют новые признаки, использующие векторы слов. Описана[39] модель word2vec, построенная на коллекции аннотаций медицинских и биологических публикаций Pubmed для идентификации похожих пар лекарство—заболевание. С помощью распределенных представлений слов выполнено[10] ранжирование возможных синонимов биомедицинских терминов из Википедии. Актуальными задачами являются создание коллекций и разработка моделей извлечения информации о заболеваниях и лекарствах для более качественного поиска гипотез о возможности перепрофилирования лекарств.

В настоящей работе высказано предположение, что на основе анализа текстов удастся создать векторное представление для лекарств, которое будет напрямую связано с их терапевтическим эффектом. Предположение основано на том, что лекарства с близкими свойствами упоминаются в подобных контекстах. Например, названия обезболивающих лекарственных препаратов фигурируют при обсуждении боли различного происхождения. Технологии семантического анализа текста позволяют создавать векторное представление слов, сохраняя их семантическое сходство и сходство контекстов их использования[41,42].

Для проверки данной гипотезы применяли комбинированный подход. На первом этапе была собрана коллекция комментариев пользователей о приеме лекарств и эффектах лечения на английском языке. Затем обучена модель word2vec, в результате работы которой каждому слову в коллекции присвоено его распределенное представление в виде вектора в евклидовом пространстве.

Проверка достоверности результатов, полученных с помощью обработки естественного языка, проведена с помощью методов хемоинформатики. Хемоинформатика является одним из основных инструментов при поиске лекарств с требуемыми свойствами[43,44]. Классическое предположение, используемое в хемоинформатике при поиске веществ с подобным биологическим действием, формулируется в виде «принципа подобия»[45]: подобные вещества обладают подобной активностью, или, более строго, вероятность найти подобные по биологической активности вещества среди структурно подобных соединений больше, чем среди структурно сильно различающихся. Эта концепция позволяет существенно сократить пространство поиска при создании новых лекарств, что исключительно полезно, если учесть объем химического пространства соединений, который по последним оценкам насчитывает до 10^{33} «лекарствоподобных» соединений[46]. Основная проблема заключается в том, что структурное подобие является исключительно сложной характеристикой и зависит от целей исследования. Классически молекулярное подобие определяется как расстояние в *D*-мерном химическом пространстве, оси координат которого образованы молекулярными дескрипторами. Молекулярный дескриптор — это заданный в численном виде инвариант молекулярного графа, отражающий тот или иной аспект молекулярной структуры. Опубликован весьма полный обзор[47] используемых в хемоинформатике дескрипторов, насчитывающих ~6000 различных типов. Однако подобие в дескрипторном пространстве не гарантирует подобия биологической активности в силу неуниверсальности дескрипторов. Набор дескрипторов, хорошо описывающих токсичность, может плохо описывать активность по отношению к белку HERG и т.д.

Нами показано, что лекарства с подобными векторами, которые получены с помощью модели word2vec, соответствуют соединениям с близким терапевтическим эффектом, в то время как химическое подобие таких лекарств невелико. Это подтверждает гипотезу, что использование инструментов анализа семантики естественного языка позволяет выявить лекарства, обладающие потенциально схожим терапевтическим эффектом и относящиеся к одной и той же терапевтической группе.

Во-первых, предложенный инструмент визуализации химического пространства лекарств на основе кластеризации семантически подобных названий будет способствовать поиску лекарственных средств, близких по использованию, что позволит более глубоко изучить терапевтические группы лекарств. Во-вторых, мы ожидаем, что данный анализ поможет находить новые вещества с потенциально подобным профилем биологической активности, что полезно при перепрофилировании лекарств.

Сбор коллекции комментариев пользователей о лекарствах

Коллекции текстов для анализа семантического подобия соединений собраны с помощью процесса сканирования веб-страниц, называемого краулинг (англ. crawling). В этом процессе используются специализированные программы — поисковые роботы, имитирующие действия пользователя при поиске информации. Поисковый робот загружает веб-страницу, читает содержимое, анализирует на наличие комментариев пользователей и копирует соответствующую информацию в базу. Комментарии пользователя определяются с помощью заданного вручную набора XPath-запросов, применяемых для доступа к частям html-документа. Поисковый робот также проходит по всем ссылкам, указанным на странице, чтобы повторить процесс сканирования. В настоящей работе использована одна из стандартных библиотек для написания поисковых роботов — Scrapy[48].

Коллекция текстов собрана со следующих популярных медицинских порталов: webmd.com[49], askapatient.com[50], drugs.com[51], dailystrength.org[52], patient.info[53]. Данные веб-ресурсы содержат дискуссионные группы, отзывы о лекарствах, ответы на вопросы, пользовательские обсуждения состояния здоровья и лечения. Дополнительно в коллекцию добавлены полученные ранее отзывы о продуктах с сайта amazon.com, которые связаны со здоровьем[54].

Начальный этап обработки собранной текстовой коллекции — токенизация. В результате этой стадии предобработки тексты разбиваются на токены: отдельные слова, цифры или знаки препинания. Статистика коллекции содержится в таблице 1. Коллекция состоит из 2.6 млн комментариев, число уникальных токенов равно 878 709. Токены, встретившиеся в коллекции менее 10 раз, из дальнейшего рассмотрения исключают как низкочастотные. Поскольку низкочастотные слова являются редкими в коллекции, представления, порожденные моделью, будут более неточными из-за небольшого количества информации о контексте с этими словами.

Таблица 1. Статистика собранных коллекций комментариев с веб-ресурсов

Ресурс	Число комментариев	Число токенов	Число уникальных токенов
webmd.com	284055	20794273	103935
patient.info	1472273	160750980	720380
drugs.com	93845	9191434	51530
amazon.com	428777	36499681	135523
askapatient.com	113836	13649150	79036
dailystrength.org	214489	13880025	76384

Векторные представления слов

В векторных представлениях слов (англ. distributed word representations, word embeddings) каждому слову отвечает вектор из вещественных чисел. Векторные представления слов основываются на предположении, что геометрические соотношения в этом пространстве будут соответствовать семантическим соотношениям между словами. Например, ближайшие соседи слова окажутся его синонимами или другими тесно связанными по смыслу словами.

В классических моделях машинного обучения начальное представление текста, вход модели, состоит из слов, закодированных в виде так называемого представления «one-hot»[55]: каждое слово соответствует одному элементу входного вектора. Таким образом, каждое слово в представлении «one-hot» является вектором из нулей с одной единицей в позиции, отвечающей данному слову. Модели обучают на данных векторах, и с помощью метода снижения размерности векторов на выходе каждому слову ставят в соответствие вектор более низкой размерности (англ. dimensionality reduction).

Позже была предложена[56] идея векторных представлений слов. Для этого использовали нейронную сеть с одним скрытым слоем, которая на скрытом слое обучалась предсказывать следующее слово $\mathbf{x}_i$ на основе знания некоторого количества слов слева и справа от заданного, называемого контекстом $\mathbf{x}_i,\ldots,\mathbf{x}_n$ и ставила в соответствие вектор вещественных чисел размерности d; значение d составляло несколько сотен. Данная идея легла в основу современной модели word2vec, предложенной[41,42] в виде непрерывного мешка слов (англ. continuous bag of words, CBOW). Суть модели CBOW состоит в том, чтобы научиться как можно лучше решать следующую задачу, схожую с задачей построения языковой модели: по заданному контексту слова восстановить само слово, т.е. предсказать центральное слово в окне по левому и правому контексту. Само предсказание делается моделью, очень похожей на нейронную сеть. Можно сказать, word2vec — это нейронная сеть с одним скрытым уровнем. Архитектура сети, соответствующей модели CBOW, показана на рисунке 1.

Алгоритм заключается в следующем.

• Каждый вход сети (input layer) — это вектор, состоящий из C векторов $\mathbf{x}_i$ размерности V, где V — размер словаря, а значение C определяется размером окна контекста — количеством слов, взятых слева и справа от рассматриваемого слова. Каждый вектор $\mathbf{x}_i$ соответствует слову в представлении «one-hot».

• При вычислении выхода скрытого слоя (hidden layer) берут среднее всех входных векторов $\mathbf{x}_i$; скрытый слой сети — это фактически матрица векторных представлений слов из словаря, n-я строка которой содержит представление n-го слова из словаря.

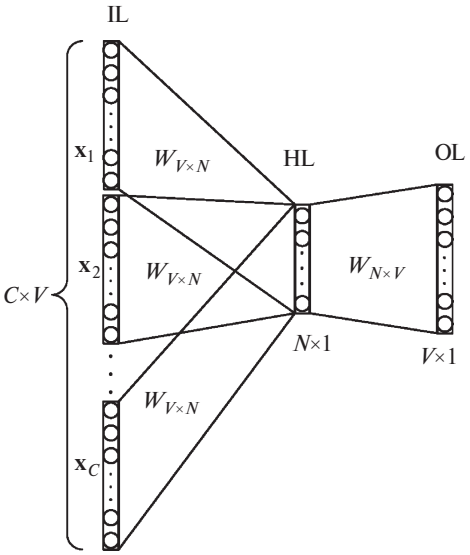


Рис. 1. Модель word2vec: CBOW (см. также лит.[57]). IL — вход сети, HL — скрытый слой, OL — выход, C — число входных векторов, определяющих размер локального контекста, $\mathbf{x}_i,...\mathbf{x}_C$ — векторы входных слов, V — размер словаря, N — число нейронов в скрытом слое, W — матрица весов соответствующего слоя сети.

• На выходе (output layer) получается вектор, состоящий из оценок вероятности p появления слов $\mathbf{x}_j$ (j = 1, ... V) из словаря. Вероятность оценивают при помощи функции softmax:

$$p(j|x_i,\dots,x_n)=\frac{\exp(x_j)}{\sum_{j=1}^{V}\exp(x_j)}.$$

В данной работе использована одна из стандартных реализаций word2vec в библиотеке Gensim[58]. Мы обучили модель CBOW со следующими параметрами: словарь из 93 526 токенов, размер окна локального контекста (определяет число векторов, подаваемых на вход) равен десяти, а размерность выходного вектора — 200.

Анализ подобия лекарственных средств

Для анализа подобия лекарственных средств рассматривают две метрики:

1) косинусная мера сходства (англ. cosine similarity) — традиционно используемая мера оценки семантического подобия слов, представленных в виде векторов;

2) коэффициент Танимото (Жаккарда), характеризующий степень подобия двух дескрипторных векторов соединений.

Косинусная мера сходства. Для каждого вектора $\mathbf{x}_i$ и $\mathbf{x}_j$ из модели word2vec косинусную меру COS рассчитывают как нормированное на длину векторов скалярное произведение, она принимает значения от –1 до 1:

$$\mathrm{COS}(x_i,x_j)=\frac{x_i x_j}{\|x_i\|\|x_j\|}.$$

Подобие химических соединений. Для расчета подобия химических соединений нами использован коэффициент Танимото (Жаккарда). Для его вычисления соединения представляют бинарными векторами, в которых единица соответствует наличию какого-либо фрагмента в молекуле, а нуль — его отсутствию. Коэффициент Танимото (TNM) рассчитывают по следующей формуле:

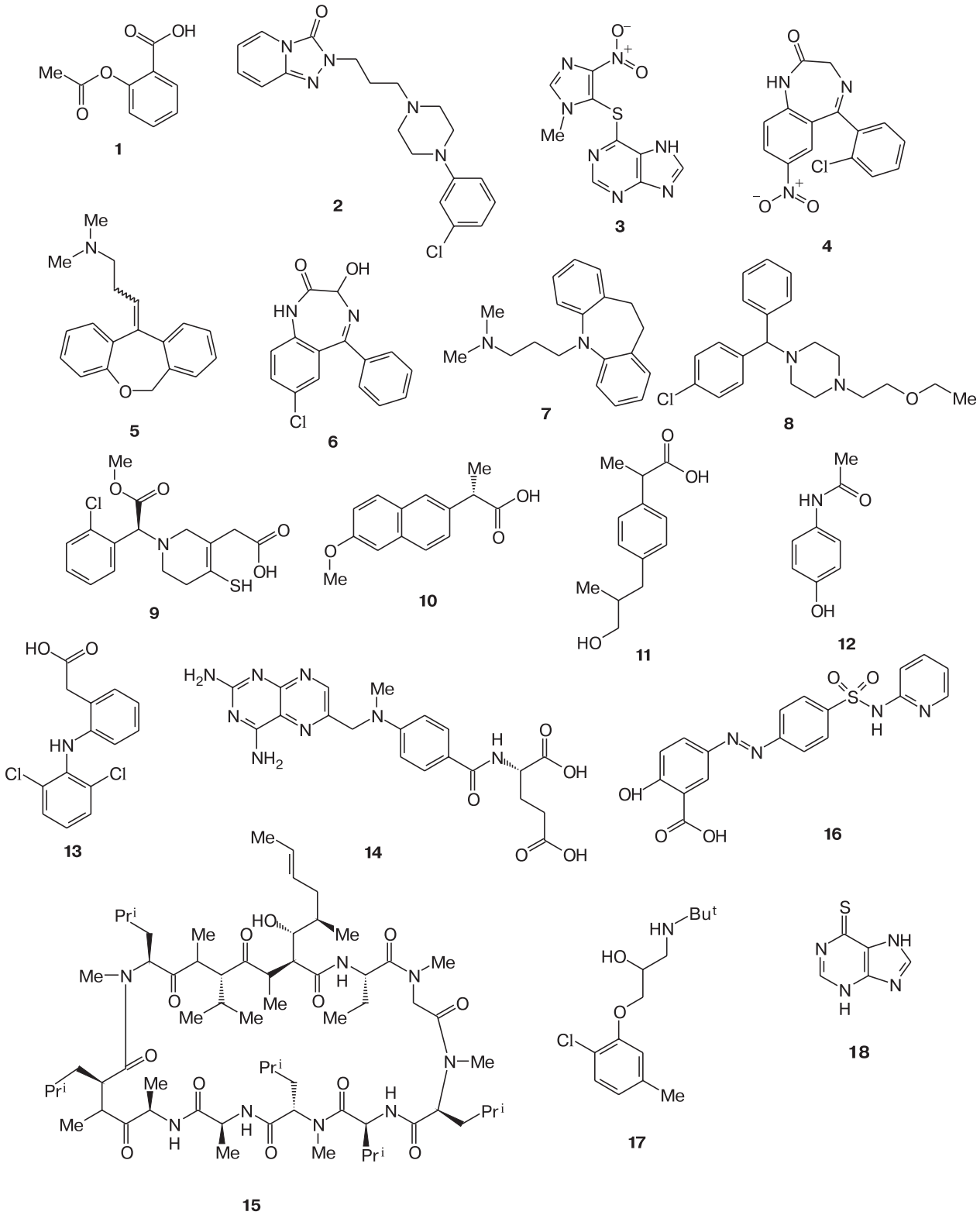
$$\mathrm{TNM}=\frac{N_C}{N_A+N_B+N_C},$$

где N_A и N_B — число уникальных признаков для каждой молекулы A и B соответственно, N_C — число общих признаков молекул A и B. Коэффициент Танимото варьируется от 0 до 1, где 1 означает, что два объекта идентичны, а 0 — объекты совершенно разные. Коэффициент Танимото является самой распространенной мерой химического подобия, применяемой в химических базах данных и моделировании.

Бинарные векторы соединений A и B рассчитывали с учетом векторов фрагментных дескрипторов ISIDA, полученных с помощью программного комплекса ISIDA Fragmentor[59]. При создании битовой строки информацией о частоте встречаемости фрагментов пренебрегали. При генерации бинарного вектора в качестве фрагментов служили так называемые расширенные атомы[59] — корневые графы-деревья. Последние представляют собой не содержащие циклов графы, у которых одна вершина (корень) помечена, а также приведены все вершины и связи, удаленные от корней на заданное топологическое расстояние (число связей). В данной работе использовались расширенные атомы диаметром от двух до четырех атомов, при этом при кодировании фрагмента принимали во внимание типы химических связей (ординарная, двойная, тройная, ароматическая) и заряды на атомах (0, –1, +1 и т.д.).

Обсуждение полученных результатов

В результате проведенного исследования установлено, что векторы, полученные семантическим анализом текстов, гораздо лучше отражают биологическую активность лекарств, чем фрагментные дескрипторы. Для примера рассмотрим три лекарства: аспирин (**1**, известный анальгезирующий, жаропонижающий, противовоспалительный и антиагрегантный лекарственный препарат), антидепрессант тразодон (**2**) и иммунодепрессант азатиоприн (**3**). В таблице 2 приведен список ближайших соседей лекарств, обнаруженных с помощью косинусной меры подобия векторных представлений слов, т.е. тех, которые наиболее семантически подобны. Это означает, что данные лекарства встречаются в текстах отзывов в одинаковых контекстах



и замена одного препарата другим не приведет к существенному изменению общего смысла. Мы использовали названия лекарств согласно базе DrugBank[60].

Ближайшими соседями антидепрессанта тразодона (**2**) оказываются противоэпилептический препарат клоназепам (**4**), антидепрессант и анксиолитик доксепин (**5**), снотворное темазепам, антидепрессант имипрамин и «мягкий» транквилизатор гидроксизин (**8**). Опубликованы научные работы о лечении депрессии с помощью клоназепама (**4**)[61–63], поэтому его также можно отнести к группе антидепрессантов. Ближайшими соседями аспирина (**1**) являются антитромботический препарат клопидогрел (**9**), а также анальгетики напроксен (**10**), ибупрофен (**11**), ацетаминофен (**12**) и диклофенак (**13**). Наиболее похожими на иммунодепрессант и цитостатик азатиоприн (**3**) оказываются иммунодепрессанты

Таблица 2. Примеры ближайших соседей в модели word2vec

Лекарство	COS	TNM
Тразодон (**2**)		
Клоназепам (**4**)	0.768	0.198
Доксепин (**5**)	0.759	0.196
Темазепам (**6**)	0.727	0.219
Имипрамин (**7**)	0.706	0.270
Гидроксизин (**8**)	0.694	0.379
Аспирин (**1**)		
Клопидогрел (**9**)	0.694	0.336
Напроксен (**10**)	0.626	0.357
Ибупрофен (**11**)	0.604	0.259
Парацетамол (**12**)	0.589	0.287
Диклофенак (**13**)	0.545	0.196
Азатиоприн (**3**)		
Метотрексат (**14**)	0.833	0.132
Циклоспорин (**15**)	0.782	0.003
Сульфасалазин (**16**)	0.725	0.017
Бупропион (**17**)	0.546	0.126
Меркаптопурин (**18**)	0.512	0.138

метотрексат (**14**), циклоспорин (**15**) и цитостатик меркаптопурин (**18**). Сульфасалазин (**16**), как и азатиоприн (**3**), используют для лечения ревматоидного артрита, неспецифического язвенного колита, болезни Крона. Соседство азатиоприна (**3**) с антидепрессантом бупропионом (**17**) может показаться ошибкой, однако более детальный анализ данных литературы показал, что он способен понижать уровень медиатора воспаления — фактора некроза опухоли — и приводит к ремиссии болезни Крона и псориаза, так же как и азатиоприн (**3**)[64]. Из примера видно, что проведенный нами анализ дает возможность выявлять нестандартные применения лекарств и предлагать гипотезы перепрофилирования.

При этом визуальный анализ структур соединений **1—18** позволяет отметить низкое химическое подобие обсуждаемых соединений, что подтверждается оценкой молекулярного подобия на основе индекса Танимото. Последний для приведенных в таблице 2 семантически подобных групп лекарств не превышает 0.4.

Мы проанализировали корреляцию между косинусной мерой сходства и коэффициентом Танимото. Для этого было выбрано случайным образом 100 лекарств из множества наиболее часто встречающихся в тексте и вычислены косинусная мера близости и коэффициент Танимото для каждой пары лекарств (рис. 2). Как видно из рисунка 2, корреляция между семантическим и молекулярным подобием отсутствует. Коэффициент корреляции Пирсона для расчитанных значений косинусной меры и коэффициента Танимото равен всего 0.13.

Кроме того, обнаружено, что векторы, построенные на основе векторных представлений слов в тексте, подчиняются арифметическим соотношениям, отражающим их семантическое и биологическое подобие, т.е. в результате суммирования векторов названий двух лекарств получается вектор, близкий

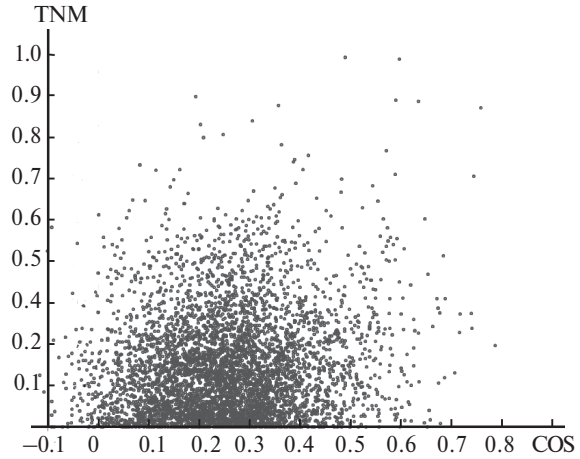


Рис. 2. Зависимость между косинусной мерой сходства (COS) и коэффициентом Танимото (TNM).

по косинусной мере к вектору лекарства, обладающего свойствами и первого, и второго препарата. Данный процесс иллюстрирует рисунок 3. Подобные соотношения встречаются и для коллекций текстов других тематик[65]. Для анализа лекарств по типу их действия использовали базу DrugBank[60]. В таблице 3 приведено несколько примеров линейных соотношений в семантическом пространстве векторов: арифметика векторов «drug#1 + drug#2» означает, что ищут слово с вектором, наиболее близким по косинусной мере сумме векторов drug#1 и drug#2.

Как видно из сравнения таблиц 2 и 3, в модели word2vec учитывают семантику текстов о лекарствах с использованием всего лишь локальных контекстов слов. Эта особенность векторов была обнаружена для обычных слов, не являющихся терминами[41,42], но она сохраняется также для векторов, соответствующих лекарствам. Рассмотрим несколько примеров, демонстрирующих данное свойство. Вектор, который получен сложением вектора, соответствующего бупропиону (**17**, антидепрессанту без седативного эффекта), и вектора темазепама (**6**, снот-

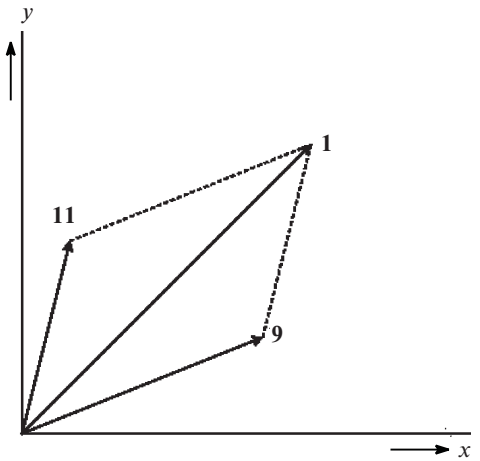
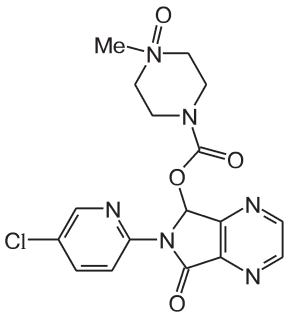
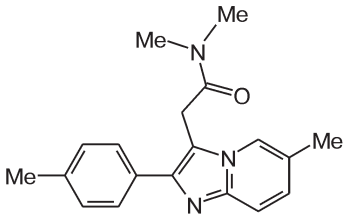


Рис. 3. Суммирование векторов слов «ибупрофен» (**11**) и «клопидогрел» (**9**); в результате получен вектор, близкий к лекарству «аспирин» (**1**).

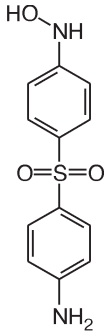
19



20



21



22

Таблица 3. Примеры линейных соотношений в модели word2vec

Лекарство	COS
Темазепан (**6**) + бупропион (**17**) = ...	
Тразодон (**2**)	0.792
Зопиклон (**20**)	0.720
Золпидем (**21**)	0.718
Клопидогрел (**9**) + ибупрофен (**11**) = ...	
Аспирин (**1**)	0.820
Напроксен (**10**)	0.806
Диклофенак (**13**)	0.768
Инфликсимаб (**19**) + + меркаптопурин (**18**) = ...	
Азатиоприн (**3**)	0.702
Дапсон (**22**)	0.638
Сульфасазолин (**16**)	0.606

ворного с седативным эффектом), наиболее близок вектору тразодона (**2**, антидепрессанту с седативным эффектом) со значением косинусной меры, равным 0.792 (см. табл. 3). Отметим, что оцененное с помощью косинусной меры подобие бупропиона (**17**) и темазепама (**6**) с тразодоном (**2**) существенно меньше. Другой интересный пример — суммирование векторов антитромботика клопидогрела (**9**) с анальгетиком ибупрофеном (**11**). На полученную сумму векторов наиболее похожим оказывается вектор аспирина (**1**), проявляющего более выраженные, чем у ибупрофена (**11**), антитромботические свойства. При этом сохраняется достаточно сильное подобие с иными негормональными противовоспалительными препаратами, которые в меньшей степени, но обладают антитромботическим эффектом. И третий проанализированный нами пример касается меркаптопурина (**18**), азатиоприна (**13**) и инфиксимаба (**19**). Меркаптопурин (**18**) характеризуется небольшим семантическим подобием с иммунодепрессантом азатиоприном (**3**), равным 0.512, поскольку меркаптопурин (**18**) обладает выраженным цитостатическим действием при меньшей активности в качестве иммунодепрессанта. В то же время меркаптопурин (**18**) является активным метаболитом азатиоприна (**3**), что отражено в их химических структурах. Инфликсимаб (**19**) — белковый иммунодепрессант, не проявляющий цитостатического действия. Сумма векторов инфликсимаба (**19**, иммунодепрессант) и меркаптопурина (**18**, цитостатик) обладает бóльшим семантическим подобием, равным 0.702, с азатиоприном (**3**, иммунодепрессант цитостатического действия), чем каждый из слагаемых индивидуально (см. табл. 2 и 3). В числе наиболее подобных векторов оказывается вектор сульфасазолина (**16**), который, как было указано выше, так же как азатиоприн (**3**), используется для лечения болезни Крона.

Таким образом, из примеров видно, что суммирование векторов лекарственных препаратов приводит к вектору лекарственного препарата, у которого присутствуют биологические свойства каждого из препаратов в отдельности. Этот эффект крайне интересен для поиска лекарств, обладающих требуемым спектром свойств. Отметим, что структурные дескрипторы не подчиняются рассматриваемым свойствам, суммирование их не приводит к значимым зависимостям такого типа.

Построение зависимостей, указанных выше, в ручном режиме может быть затруднено. Средства визуализации позволяют построить двумерную карту распределения лекарственных препаратов по семантическому подобию, которая более удобна для ручного анализа. Для визуализации были выбраны названия только тех соединений, которые относятся к категории «Нервная Система» («Nervous System») в базе DrugBank (137 соединений), а также их векторные представления. С помощью модели на основе векторных представлений слов можно кластеризовать семантически подобные слова[41,42,66–69]. В настоящей работе применена иерархическая версия алгоритма кластеризации DBSCAN (см. лит.[70]) и метод визуализации данных t-SNE (см. лит.[71]) для снижения размерности векторов из библиотеки word2map.[72] Метод кластеризации DBSCAN основан на соединении некоторых областей, плотность объектов внутри которых превышает некоторый заданный порог. На рисунке 4 представлены результаты кластеризации. Цветовая визуализация легенды отображает используемые спектры цветов и количество кластеров.

Лекарства, оказывающие влияние на нервную систему и находящиеся в одном кластере (близко

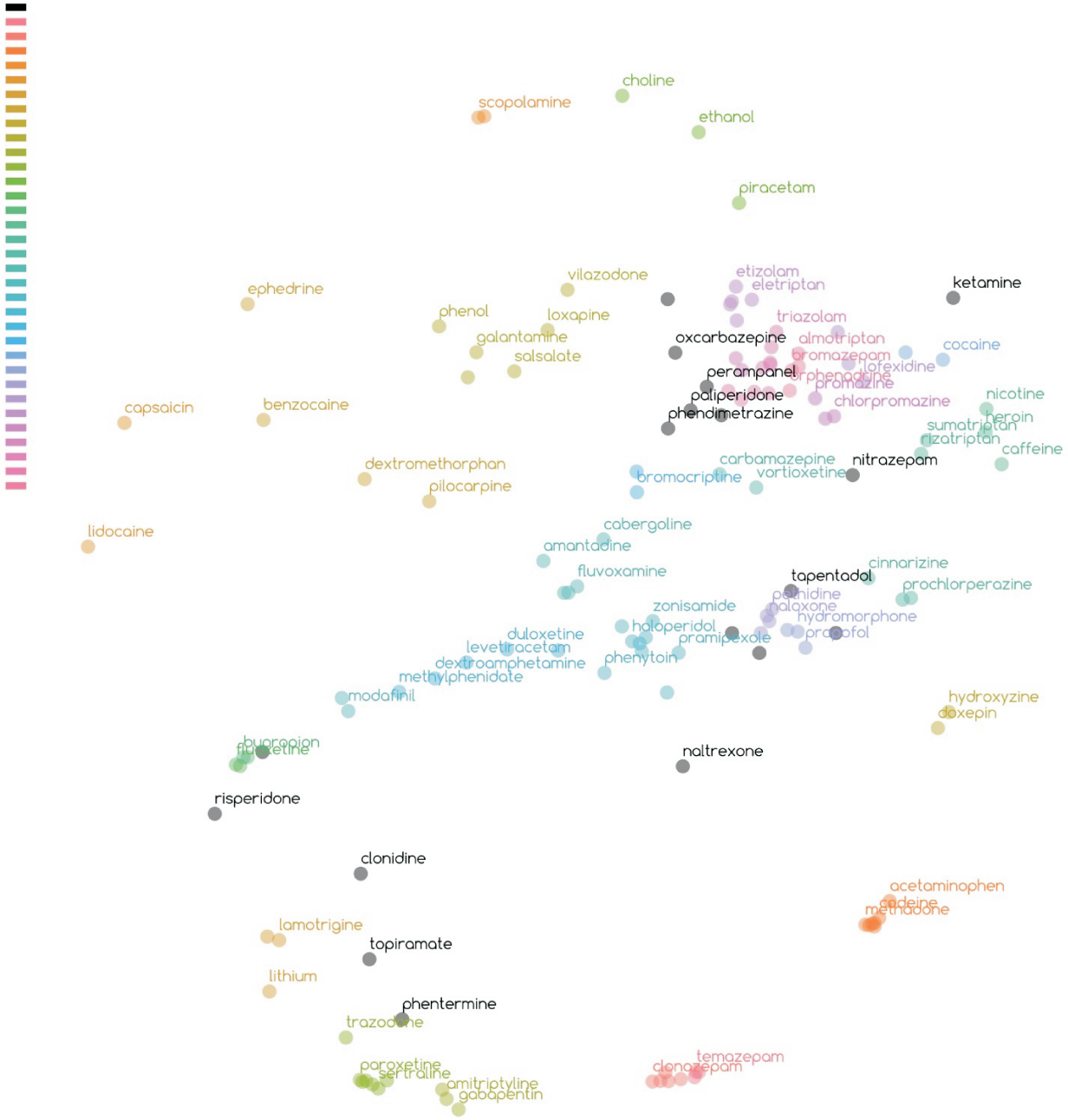


Рис. 4. Визуализация векторных представлений слов, отвечающих лекарственным средствам (часть лекарств фигурирует без названий). Цветом обозначены кластеры. Расстояние между точками соответствует расстоянию в векторном пространстве. Цветовая визуализация легенды отображает используемые спектры цветов и количество кластеров. Рядом с точками приведены названия лекарств.

в пространстве векторных представлений слов), как правило, обладают близкими биологическими действиями. Например, наркотические препараты (кокаин, героин, кетамин) и иные вызывающие привыкание вещества (кофеин, кокаин) оказываются в правой части рисунка, анестетики (эфедрин, лидокаин, бензокаин) — в левой части. Лекарственные средства группы бензодиазепинов с противосудорожным и анксиолитическим действиями клоназепам (**4**) и лоразепам находятся в одном кластере (втором по счету на легенде), который близок к темазепаму (**6**, средство из группы бензодиазепинов со снотворным действием) в другом кластере. В то же время антидепрессанты бупропион (**17**) и флуоксетин лежат в соседних кластерах зеленого оттенка. Флуоксетин — селективный ингибитор обратного захвата серотонина, бупропион (**17**) — селективный ингибитор обратного захвата норадреналина и дофамина. Это объясняется тем, что пользователи в дискуссионных группах обсуждают эффекты совместного применения обоих лекарств, что также является темой для исследований в научных статьях[73–76]. Такое свойство визуализации и кластеризации оказывается полезным при решении задачи анализа взаимодействий соединений и поиска кандидатов для перепрофилирования.

Таким образом, проведенные в настоящей работе исследования, по нашему мнению, являются доказательством возможности анализа текстов в социальных сетях для идентификации лекарств с похожими терапевтическими эффектами. С использованием методов word2vec и кластеризации получено множество кластеров соединений, содержащих лекарства с семантически подобным окружением, которые, в большинстве своем, также обладали схожим терапевтическим действием. Показано, что векторное представление слов, составленное в результате анализа естественного языка, лучше отражает биологическую активность, чем структурные дескрипторы. В будущем планируется провести более глубокий анализ этих кластеров с целью выявления лекарств с новыми терапевтическими свойствами; лекарства такого рода могут быть достаточно быстро перепрофилированы для новых применений.

Работа выполнена за счет средств субсидии, выделенной в рамках государственной поддержки Казанского (Приволжского) федерального университета в целях повышения его конкурентоспособности среди ведущих мировых научно-образовательных центров, а также для выполнения государственного задания в сфере научной деятельности (проекты № 4.1493.2017/4.6 и № 4.5151.2017/6.7) и при финансовой поддержке Российского научного фонда (проект № 15-11-10019, Е. Тутубалина, С. Николенко, работа по описанию и созданию векторных представлений слов).

Список литературы

1. E. Lekka, S. N. Deftereos, A. Persidis, A. Persidis, C. Andronis, *Drug Discovery Today: Therapeutic Strategies*, 2012, **8**, 103.
2. W. Loging, R. Rodriguez-Esteban, J. Hill, T. Freeman, J. Miglietta, *Drug Discovery Today: Therapeutic Strategies*, 2012, **8**, 109.
3. N. C. Baker, B. M. Hemminger, *J. Biomed. Inform.*, 2010, **43**, 510.
4. R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, *Proc. 2010 Workshop on Biomedical Natural Language Processing* (*Uppsala, Sweden, July 15, 2010*), Uppsala, 2010, p. 117.
5. A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. E. Leonard, J. H. Holmes, *J. Biomed. Inform.*, 2011, **44**, 989.
6. C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, N. Dasgupta, *Drug Safety*, 2014, **37**, 343.
7. A. Nikfarjam, A. Sarker, K. O´Connor, R. Ginn, G. Gonzalez, *J. Am. Med. Inform. Ass.*, 2015, 1.
8. S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, C. Paris, *ACM Computing Surveys*, 2015, **47**, 56.
9. C. C. Huang, Z. Lu, *Brief. Bioinform.*, 2016, **17**, 132.
10. C. H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, C.W. Thomas, Z. Lu, *Proc. 5th BioCreative Challenge Evaluation Workshop*, 2015, 154.
11. M. Rastegar-Mojarad, H. Liu, P. Nambisan, *JMIR Res. Protocols*, 2016, **5**.
12. A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, G. Gonzalez, *J. Biomed. Inform.*, 2015, **54**, 202.
13. J. Lardon, R. Abdellaoui, F. Bellet, H. Asfari, J. Souvignet, N. Texier, *J. Med. Internet Res.*, 2015, **17**, 171.
14. H. J. Murff, V. L. Patel, G. Hripcsak, D. W. Bates, *J. Biomed. Inform.*, 2003, **36**, 131.
15. R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, N. H. Shah, *J. Drug Safety*, 2014, **37**, 777.
16. R. Sloane, O. Osanlou, D. Lewis, D. Bollegala, S. Maskell, M. Pirmohamed, *British J. Clin. Pharmacol.*, 2015, **80**, 910.
17. A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. H. Leonard, J. H. Holmes, *J. Biomed. Inform.*, 2011, **44**, 989.
18. C. C. Yang, H. Yang, L. Jiang, M. Zhang, *Proc. 2012 Intern. Workshop on Smart Health and Wellbeing* (*Sheralon, October 29, 2012*), Sheralon, 2012, 33.
19. X. Liu, H. Chen, *Proc. Intern. Conf. Smart Health* (*Beijing, August 3—4, 2013*), Beijing, 2013, 134.
20. S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep, R. Srinivasan, *J. BMC Med. Inform. Decisicion Making*, 2014, **14**.
21. C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, N. Dasgupta, *J. Drug Safety*, 2014, **37**, 343.
22. K. O´Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, G. Gonzalez, *Proc. Am. Med. Inform. Association (AMIA) Ann. Symp.*, 2014, **2014**, 924.
23. C. C. Yang, H. Yang, L. Jiang, *J. ACM Trans. Management Inform. Systems*, 2014, **5**, 2.
24. E. Tutubalina, S. Nikolenko, *J. Computaciyn y Sistemas*, **2017**, 21.
25. J. C. Na, W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y. K. Chang, Y. L. Theng, *Proc. Intern. Conf. on Asian Digital Libraries* (*Taivan, November 12—15, 2012*), Taivan, 2012, 189.
26. A. Nikfarjam, G. H. Gonzalez, *Proc. AMIA Ann. Symp. (Washington, October 22—26, 2011)*, Washington, 2011, **2011**, p. 1019.
27. Y. Niu, X. Zhu, J. Li, G. Hirst, *J. AMIA*, 2005, **2005**, 507.
28. J. Bian, U. Topaloglu, F. Yu, *Proc. 2012 Intern. Workshop on Smart Health and Wellbeing* (*Sheralon, October 29, 2012*), Sheralon, 2012, 25.
29. M. Yang, X. Wang, M. Y. Kiang, *PACIS*, 2013, 193.
30. A. Patki, A.Sarker, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. O´Connor, K. Smith, G. Gonzalez, *Proc. BioLinkSig 2014 (Boston, July 11—12, 2014)*, Boston, 2014, **2014**, p. 1—8.
31. B. W. Chee, R. Berlin, B. Schatz, *Proc. AMIA Ann. Symp. (Washington, October 22—26, 2011)*, Washington, 2011, **2011**, 217.
32. A. Sarker, R. Ginn, A. Nikfarjam, K. O´Connor, K. Smith, S. Jayaraman, U. Tejaswi, G. Gonzalez, *J. Biomed. Inform.*, 2015, **54**, 202.
33. R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, G. Gonzalez, *Proc. 2010 Workshop on Biomedical Natural Language Proc.*, 2010, 117—125.
34. A. Yates, N. Goharian, O. Frieder, *Proc. 2013 ACM SIGIR Workshop on Health Search and Discovery (Dublin, Ireland, August 1, 2013)*, 2013, p. 55.
35. E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, K. Ohe, *J. Stud. Health. Technol. Inform.*, 2010, **160**, 739.
36. M. A. J. I. D. Rastegar-Mojarad, R. K. Elayavilli, Y. Yu, H. Liu, *Proc. Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing*, Big Island of Hawaii, 2016; http://diego.asu.edu/psb2016/acceptedpapers/Mayo-NLP.pdf.

37. T. Huynh, Y. He, A. Willis, S. Rüger, *Proc. COLING 2016, 26th Intern. Conf. on Computational Linguistics: Technical Papers (Osaka, December 11—17, 2016)*, Osaka, 2016, 877.
38. Y. Wu, J. Xu, M. Jiang, Y. Zhang, H. Xu, *AMIA Ann. Symp. Proc.*, 2015, 1326.
39. D. L. Ngo, N. Yamamoto, V. A. Tran, N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. Kubo, K. Satou, *J. Biomed. Sci. Eng.*, 2016, **9**, 7.
40. A. N. Jagannatha, J. Chen, H. Yu, *Proc. 6th Intern. Workshop on Health Text Mining and Information Analysis (Louhi, 2015)*, Louhi, 2015, 142.
41. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, *Proc. of NIPS*, Eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, Online, 2013, 3111.
42. T. Mikolov, K. Chen, G. Corrado, J. Dean, *arXiv preprint arXiv*, 2013.
43. B. F. Begam, J. S. Kumar, *Proc. Eng.*, 2012, **38**, 1264.
44. A. Varnek, I. I. Baskin, *Mol. Inform.*, 2011, **30**, 20.
45. M. A. Johnson, G. M. Maggiora, in *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, Hoboken, New Jersey, 1990, p. 394.
46. P. G. Polishchuk, T. I. Madzhidov, A. Varnek, *J. Computer-Aided Molecular Design*, 2013, **27**, 675.
47. R. Todeschini, V. Consonni, *Molecular Descriptors for chemoinformatics*, John Wiley & Sons, Hoboken, New Jersey, 2009.
48. URL: https://scrapy.org.
49. URL: http://www.webmd.com.
50. URL: http://www.askapatient.co.
51. URL: https://www.drugs.com.
52. URL: https://dailystrength.org.
53. URL: http://patient.info.
54. J. McAuley, C. Targett, Q. Shi, Van Den Hengel, *Proc. 38th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, ACM, New York, 2015, p. 43—52.
55. J. Beck, B. Woolf, *J. Intelligent Tutoring Systems*, 2000, 584.
56. Y. Bengio, R. Ducharme, P. Vincent, *J. Machine Learning Res.*, 2003, **3**, 1137.
57. X. Rong, *arXiv preprint arXiv*, 2014; https://arxiv.org/pdf/1411.2738.
58. R. Rehurek, P. Sojka, *Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta, Malta, May 22, 2010)*, ELRA, 2010, p 45.
59. A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov´ev, *J. Comput. Aided. Mol. Des*, 2005, **19**, 693.
60. https://www.drugbank.ca/.
61. A. Kishimoto, K. Kamata, T. Sugihara, S. Ishiguro, H. Hazama, R. Mizukawa, N. Kunimoto, *Acta Psychiatrica Scandinavica*, 1988, **77**, 81.
62. S. Morishita, S. Aoki, *J. Affective Disorders*, 1999, **53**, 275.
63. B. W. Dunlop, P. G. Davis, *Prim Care Companion J. Clin. Psychiatry*, 2008, **10**, 222.
64. S. V. Kane, E. L. Altschuler, R. E. Kast, *Gastroenterology*, 2003, **125**, 1290.
65. T. Mikolov, W. Yih, G. Zweig, *Proc. 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Atlanta, 2013)*, 2013, **13**, 746.
66. E. Tutubalina, S. Nikolenko, *Proc. Intern. Conf. on Analysis of Images, Social Networks and Texts (Екатеринбург, 2016)*, *Бизнес-центр Палладиум*, Екатеринбург, 2016, с. 208.
67. S. I. Nikolenko, *Proc. 39th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval (Pisa, 2016)*, Pisa, 2016, p. 1029.
68. N. A. Loukachevitch, *Proc. Intern. Conf. on Text, Speech, and Dialogue (Москва, 2016)*, РГГУ, Москва, 2016, с. 134.
69. V. Solovyev, V. Ivanov, *J. Comput. Intelligence and Neurosci.*, 2016; doi: 10.1155/2016/4183760.
70. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *Proc. Second Intern. Conf. on Knowledge Discovery and Data Mining (Port Land, 1996)*, AAAI, Menlo Park, 1996, p. 226.
71. L. V. D. Maaten, G. Hinton, *J. Machine Learning Res.*, 2008, **9**, 2579.
72. L. van der Maaten, *J. Machine Learning Res.*, 2008, **9**, 2579.
73. S. X. M. Li, K. W. Perry, D. T. Wong, *Neuropharmacology*, 2002, **42**, 181.
74. J. A. Bodkin, R. A. Lasser, Jr., J. D. Wines, D. M. Gardner, R. J. Baldessarini, *J. Clinical Psychiatry*, 1997, **58**, 137.
75. P. Blier, H. E. Ward, P. Tremblay, L. Laberge, C. Hébert, R. Bergeron, *Am. J. Psychiatry*, 2009, **167**, 281.
76. З. Ш. Мифтахутдинов, Е. В. Тутубалина, А. Э. Тропша, *Компьютер. лингвистика и интеллектуальные технол.*, 2017, **1**, 155.

Поступила в редакцию 3 апреля 2017;
после доработки — 29 сентября 2017