УДК 004.912, 004.89

ОБЗОР МЕТОДОВ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Белов Сергей Дмитриевич[1], Зрелова Дарья Петровна[2], Зрелов Петр Валентинович[3] Кореньков Владимир Васильевич[4]

[1]Ведущий программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
Российский экономический университет имени Г.В. Плеханова;
117997, РФ, г. Москва, Стремянный пер., 36;
e-mail:belov@jinr.ru.

[2]Инженер-программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: zrelova@me.com.

[3]Начальник отдела, кандидат физико-математических наук;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
Российский экономический университет имени Г.В. Плеханова;
117997, РФ, г. Москва, Стремянный пер., 36;
e-mail: *zrelov@jinr.ru.*

[4]Директор, доктор технических наук, профессор;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
ГБОУ ВО МО «Университет «Дубна»,
Институт системного анализа и управления;
141980, Московская обл., г. Дубна, ул. Университетская, 19;
Российский экономический университет имени Г.В. Плеханова;
117997, РФ, г.Москва, Стремянный пер., 36;
e-mail: korenkov@cv.jinr.ru.

В работе представлен краткий обзор современных методов и подходов, используемых для автоматической обработки текстовой информации. В англоязычной литературе эта область науки носит название NLP – Natural Language Processing. Само название говорит о том, что предметом анализа (а для многих задач – и синтеза) являются материалы, изложенные на одном из естественных языков (а для ряда задач – на нескольких языках одновременно), т.е. национальных языках общения между людьми. Языки программирования в эту группу не входят. В русскоязычной литературе эта область носит название компьютерной (или математической) лингвистики. Обычно NLP (компьютерная лингвистика) наряду с анализом текста включает также анализ речи, но в данном обзоре последнее не рассматривается.

При составлении обзора использовались материалы оригинальных работ, монографий и ряда статей журнала «Открытые системы.СУБД».

Ключевые слова: NLP, лингвистика, корпус языка, машинное обучение, модель Word2vec, нейронные сети.

Для цитирования:
Обзор методов автоматической обработки текстов на естественном языке / С. Д. Белов, Д. П. Зрелова, П. В. Зрелов, В. В. Кореньков // Системный анализ в науке и образовании: сетевое научное издание. – 2020. – № 3. – С. 8–22. – URL: http://sanse.ru/download/401.

OVERVIEW OF METHODS FOR AUTOMATIC NATURAL LANGUAGE TEXT PROCESSING

Belov Serey[1], Zrelova Daria[2], Zrelov Petr[3], Korenkov Vladimir[4]

[1]Lead programmer;
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Plekhanov Russian University of Economics;
36 Stremyanny per., Moscow, 117997, Russia;
e-mail: belov@.jinr.ru.

[2]Programmer engineer;
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
e-mail: zrelova@me.com.

[3]Head of Department, Candidate of Physics and Mathematics;
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Plekhanov Russian University of Economics;
36 Stremyanny per., Moscow, 117997, Russia;
e-mail: zrelov@jinr.ru.

[4]Director, Doctor of Technical Science, professor;
Joint institute for nuclear research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, Joliot-Curie, 6;
Dubna State University;
Institute of the system analysis and management;
141980, Dubna, Moscow reg., Universitetskaya str., 19;
Plekhanov Russian University of Economics;
36 Stremyanny per., Moscow, 117997, Russia;
e-mail: korenkov@cv.jinr.ru.

This paper provides a brief overview of modern methods and approaches used for automatic processing of text information. In English-language literature, this area of science is called NLP-Natural Language Processing. The very name suggests that the subject of analysis (and for many tasks – and synthesis) are materials presented in one of the natural languages (and for a number of tasks – in several languages simultaneously), i.e. national languages of communication between people. Programming languages are not included in this group. In Russian-language literature, this area is called Computer (or mathematical) linguistics. NLP (computational linguistics) usually includes speech analysis along with text analysis, but in this review speech analysis does not consider.

The review used materials from original works, monographs, and a number of articles published the «Open Systems.DBMS» journal.

Keywords: NLP, linguistics, language corpus, machine learning, Word2vec model, neural networks.

For citation:
Overview of methods for automatic natural language text processing = Обзор методов автоматической обработки текстов на естественном языке / S. Belov, D. Zrelova, P. Zrelov, V. Korenkov // System Analysis in Science and Education – 2020. – № 3. – Pp. 8–22. – URL: http://sanse.ru/download/401.

Введение

Обработка естественного языка – это область исследований в компьютерных науках и искусственном интеллекте (ИИ). Обработка обычно включает в себя перевод естественного языка в числовые данные, с помощью которых компьютер может получить информацию об окружающем мире. Для этих целей разрабатываются так называемые технологии NLP (Natural Language Processing). В 2010-х годах обработка естественного языка и диалоговые машины (чат-боты) на основе NLP начали получать все большее распространение. Сначала поиск Google напоминал лишь работу с предметным указателем – инструментом, для использования которого не требовалось особых навыков. Но скоро он стал интеллектуальнее и начал понимать поисковые запросы, близкие к естественному языку. Затем появилась автодополнения в смартфонах. С появлением ботов, подобных боту Tau от Microsoft стало ясно, что NLP-боты влияют на социум. Боты стали собирать твиты, чтобы прогнозировать результаты выборов, а позднее – влиять на эти результаты. Появились системы для предсказания экономических трендов, подобные алгоритмы стали все больше влиять на экономику и привели к сдвигу в общественном сознании –люди стали применять NLP для усиления роли «машин» при принятии собственных решений. Благодаря бурному потоку неструктурированных данных по вопросам политики и экономики NLP становился неотъемлемым инструментом политтехнологов и финансистов. Генерация все большего объема развлекательного, рекламного и финансово-отчетного контента не требует участия человека. Компьютерные игры и виртуальные миры содержат NLP-боты, способные общаться с человеком. NLP обеспечивает эффективный информационный поиск и с помощью механизмов фильтрации или продвижения определенных страниц влияет на потребляемую пользователем информацию. Поиск – исторически первая коммерчески успешная сфера приложения NLP. Поиск вдохновлял все более быструю разработку NLP-алгоритмов, которые затем усовершенствовали технологии поиска на основе методик индексации и прогнозирования NLP.

Поисковая машина может предоставить более точные результаты поиска, если индексирует вебстраницы или архивы документов таким образом, что принимается во внимание значение. Автозаполнение, использующее NLP, получило широкое распространение в поисковых системах и клавиатурах мобильных устройств. Во множестве текстовых процессоров, плагинов для браузера и текстовых редакторов есть встроенные средства проверки орфографии, грамматики и согласования слов. Некоторые из диалоговых машин (чат-ботов) используют поиск на естественном языке, чтобы найти ответ на сообщение своего собеседника. Системы NLP могут использоваться для создания не только коротких ответов в диалоговых машинах, виртуальных помощниках, генерации коротких постов в социальных сетях, но и для составления более длинных отрывков текста. The Associated Press использует роботов-журналистов на базе NLP для написания целых статей по тематике финансов и репортажей со спортивных состязаний [1]. Спам-фильтры на базе NLP в ранних программах электронной почты способствовали тому, что в 1990-е годы электронная почта обогнала телефон и факс в качестве канала связи. Примерно 20 % твитов о президентских выборах в США в 2016 году были составлены чат-ботами [2]. Эти боты поддерживают выгодные их владельцам и разработчикам точки зрения. Технологии NLP используются для составления обзоров фильмов и отзывов на товары из магазинов. Большее количество отзывов – продукт работы автономных конвейеров NLP, которые никогда не были в кинотеатре или не покупали обозреваемый продукт.

Другие задачи, решаемые с помощью прикладных NLP-систем – машинный перевод, реферирование, аннотирование и анализ тональности текста, рубрицирование, классификация и кластеризация текста, построение баз знаний и др.

1. Лингвистические ресурсы

Разработка и применение NLP-систем опирается на использование тех или иных лингвистических ресурсов: лексических (словарных) и текстовых. К лексическим ресурсам относятся словари, тезаурусы и онтологии.

Словари являются наиболее традиционной формой представления лексической информации; они различаются своими единицами (обычно слова или словосочетания), структурой, охватом лексики (словари терминов конкретной проблемной области, словари общей лексики, словари синонимов или паронимов и т.п.). Единица словаря называется словарной статьей, в ней представляется информация о лексеме. Лексические омонимы обычно представляются в разных словарных статьях. К лексическим

ресурсам относятся базы словосочетаний, в которые отбираются наиболее типичные словосочетания конкретного языка. Более сложными видами лексических ресурсов являются тезаурусы и онтологии. Тезаурус – это семантический словарь, т.е. словарь, в котором представлены смысловые связи слов – синонимические, отношения «род-вид», «часть-целое», ассоциации. В качестве характерного примера можно привести информационно-поисковый тезаурус РуТез [3] для русского языка, охватывающего общественно-политическую лексику. С понятием тезауруса тесно связано понятие онтологии. Онтология – набор понятий и сущностей в определенной области знаний, ориентированный на многократное использование для различных задач. Онтологии могут создаваться на базе существующей лексики – в этом случае они называются лингвистическими, например, WordNet [4] – большой лексический ресурс, в котором собраны слова английского языка: существительные, прилагательные, глаголы, наречия, их смысловые связи. По схеме английского WordNet были построены аналогичные лексические ресурсы для других европейских языков, объединённые под общим названием EuroWordNet. Текстовые ресурсы, служащие для построения модулей лингвистических процессоров, охватывают коллекции текстов (обычно для конкретных проблемных областей) и текстовые корпуса. Корпус текстов – это представительный массив текстов, собранный по определённому принципу (по жанру, авторской принадлежности и т.п.) и обладающий лингвистической разметкой – морфологической, синтаксической, дискурсивной и др. В России наиболее известными являются Национальный корпус русского языка (НКРЯ) [5] и ГИКРЯ [6], они отличаются целями и методами создания, набором включенных русскоязычных текстов. Размеченные корпуса используются для лингвистических исследований и обучения лингвистических процессоров на основе методов машинного обучения [7].

2. Типы NLP-систем

В настоящее время существует два главных подхода: основанный на правилах (rule-based), или инженерный, и основанный на машинном обучении (machine learning). Исторически первым является подход на правилах, который заключается в описании необходимой лингвистической информации в виде формальных правил. Правила создаются лингвистами или специалистами-прикладниками в конкретной проблемной области. В подходе, основанном на машинном обучении, источником информации выступают не правила, а отобранные тексты проблемной области. Среди методов, применяемых в рамках подхода, выделяют методы обучения с учителем (supervised), методы обучения без учителя (unsupervised) и методы частичного обучения с учителем (bootstrapping). Чаще всего применяется обучение с учителем, при котором происходит построение математической и программной модели – машинного классификатора, который умеет распознавать различные классы единиц текста (слов, словосочетаний и других конструкций) или самих текстов. Построение классификатора происходит на специально размеченном текстовом корпусе (обучающей выборке), в котором распознаваемым единицам (или самим текстам) приписаны метки, кодирующие важные признаки распознаваемых единиц/текстов. Обучение представляет собой выявление общих закономерностей, присущих текстам.

Оба рассмотренных подхода имеют свои достоинства и недостатки. Создание правил трудоемко и требует достаточно квалифицированного труда лингвиста. В то же время правила обычно декларативны и легко понимаемы, поэтому их просто модифицировать и расширять. Машинное обучение не требует ручного труда по составлению правил и сокращает время разработки систем, однако необходимы знания для выбора подходящих методов обучения. Кроме того, результирующие модели (классификаторы) не имеют явной лингвистической интерпретации. Также машинное обучение предполагает наличие подходящего размеченного корпуса текстов, что не всегда возможно. Современная тенденция – создание модульных систем автоматической обработки текстов, в которых разные модули могут быть созданы в рамках разных подходов, например, модуль графематического анализа – на основе машинного обучения, а морфологического – на основе правил. Машинное обучение довольно часто применяется для обработки коллекций текстовых документов, с использованием признаковой модели текста, при которой признаки определены для каждого документа по отдельности. Признаками могут выступать различные информационные характеристики текста: как лингвистические, так статистические и структурные: например, частота определенных слов (или их категорий) в документе, частота использования спецзнаков, соотношение частей речи слов, наличие определенных синтаксических конструкций или разделов текста, дата создания и др.

Разновидностями признаковой модели являются модель BOW (bag of words – мешок слов), в которой текст характеризуется набором своих значимых слов (обычно это все знаменательные слова,

точнее, их леммы), а также векторная модель текста, в которой указанный набор упорядочен. Векторная модель применяется, например, в информационном поиске, при этом в качестве признаков чаще берутся не слова, а более сложные характеристики, такие как показатель TF-IDF [8] для слов. Отдельно рассматривается статистическая языковая модель (Language Model), которая строится по представительному массиву текстов путем подсчета частот стоящих рядом слов (N-грамм). Чаще всего рассматриваются биграммы (N = 2) и триграммы (N = 3). Модель отвечает на вопрос насколько вероятно появление заданного слова, если непосредственно перед ним встречались определенные слова. Вероятности рассчитываются на основе собранной статистики. Разновидностями модели являются N-граммы частей речи слов текста или N-граммы букв текста (возможны и другие модели) [7].

Классическая схема анализа текста

Самые большие возможности и высокое качество анализа текстов можно получить, проведя его полный анализ. Однако сложности, возникающие при создании подобного анализа таковы, что на практике до сих пор не реализованы все теоретические положения, разработанные на данный момент. Основными проблемами здесь являются сложность синтаксического анализа текста и сложность создания полноценной экспертной системы. Для полноценной работы система анализа текста должна иметь возможность проанализировать тест, поданный пользователем на вход, с точки зрения синтаксиса (структуры предложений), семантики (понятий, применяемых в тексте) и прагматики (правильности употребления понятий и целей их употребления). Далее система должна сгенерировать свой отклик во внутреннем представлении, пригодном для логического вывода, и синтезировать свой отклик на естественном языке. В целом система, поддерживающая полный анализ, должна содержать в себе следующие модули. Графематический анализ – обеспечивает выделение синтаксических или структурных единиц из входного текста, который может представлять собой линейную структуру, содержащую единый фрагмент текста. Морфологический анализ – обеспечивает определение нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе. Синтаксический анализ – самая сложная часть анализа текста. Здесь необходимо определить роли слов и их связи между собой. Результатом этого этапа является набор деревьев, показывающих такие связи. Семантический анализ проводит анализ текста «по смыслу». С одной стороны, семантический анализ уточняет связи, которые не смог уточнить постсинтаксический анализ, так как многие роли выражаются не только при помощи средств языка, но и с учетом значения слова. С другой стороны, семантический анализ позволяет отфильтровать некоторые значения слов или даже целые варианты разбора как «семантически несвязные». Этапом семантического анализа заканчивается анализ входного текста. В случае необходимости генерации отклика, например, в ходе диалога с пользователем или при переводе документов с иностранного языка к рассмотренным этапам добавляются этапы синтеза – синтаксический, морфологический и графематический. Генерация отклика в разной мере присуща всем видам диалоговых систем, некоторым видам систем составления рефератов текста, статистического анализа текста, генерации текстов. Отклик выбирается из некоторого корпуса текстов или генерироваться «на лету» [9].

Ограничения существующих методов

Обработка текста происходит в несколько этапов, при этом выходные данные одного этапа передаются либо на вход следующего этапа, либо модулям, предназначенным для решения конкретной практической задачи, например, машинного перевода. Результаты работы любого этапа неоднозначны – одни и те же входные данные дают несколько возможных результатов, и наоборот – один и тот же результат может получаться из совершенно разных входных данных. Это приводит к тому, что объем обрабатываемой информации при переходе с этапа на этап может резко увеличиваться.

Традиционная схема многоэтапной работы требует пересмотра – анализ текста может проходить в две фазы: поверхностное сканирование и выявление ключевых фрагментов текста, а затем детальный анализ по принципиально другому сценарию, согласно которому вышестоящие модули обращаются к нижестоящим за необходимой уточняющей информацией. Напри-

мер, в задаче извлечения информации при поверхностном сканировании определяются упоминания людей, а затем семантический модуль обращается к нижележащему синтаксическому с запросом «какие связи есть у каждого найденного упоминания» [9].

Некоторые подходы к смысловому разбору текстов

Латентно-семантический анализ текста *(Latent semantic analysis, LSA).* Латентно-семантический анализ [10] отображает документы и отдельные слова в так называемое «семантическое пространство», в котором производятся все дальнейшие действия. При этом делаются следующие предположения: документы – это набор слов, порядок слов в документах игнорируется, важна только частота встречаемости слова в документе; семантическое значение документа определяется набором слов, которые, как правило, употребляются вместе. Например, в биржевых сводках часто встречаются слова: «фонд», «акция», «доллар»; каждое слово имеет единственное значение.

TF-IDF анализ текста. TF-IDF анализ является одним из методов машинного обучения. TF-IDF – это статистический показатель, который используется преимущественно для оценки конкретного слова в контексте всего документа, входящего в общую коллекцию. Термин TF/IDF имеет англоязычное происхождение, где TF дословно означает частотность вхождения термина (term frequency), а IDF – инвертированная частота документа (inverse document frequency). В соответствии с отношением TF/IDF весомость определенного слова прямо зависит от того, сколько раз оно встречается в конкретном тексте и обратно зависима от числа использования данного слова во множестве остальных документов. TF или частота слова – это отношение количества вхождения конкретного термина к суммарному набору слов в исследуемом документе. Этот показать отражает важность слова в рамках определенной статьи/публикации. IDF или инвертированная частота документа – это обратное значение частоты, с которой определенное слово фигурирует в коллекции документов. Благодаря данному показателю можно снизить весомость наиболее широко используемых слов (предлогов, союзов, общих терминов и понятий). Для каждого термина в рамках определенной базы текстов предусматривается лишь одно единственное значение IDF. Если слово встречается в каком-либо документе часто, а во всех остальных документах – редко, то это слово имеет большую значимость для того самого документа. Стоит также отметить реализацию метода TF-IDF в библиотеке skLearn [11], простую в реализации и обладающую множеством полезных входных параметров.

Векторное представление слов. Главная идея векторной модели семантики (vector space model, VSM) – это представление каждого текста коллекции в качестве точки в многомерном пространстве (вектора в векторном пространстве). Близко лежащие точки соответствуют семантически схожим документам. Для представления слов в векторном формате существуют уже готовые методы, одним из самых популярных является технология Word2Vec [12, 13]. В 2013 году Томаш Миколов опубликовал работу [12], в которой предлагал использовать нейронные сети для обучения векторам слов, но для меньшей размерности: по кортежам (слово, контексты) обучалась нейронная сеть простейшей архитектуры, на выходе каждому слову в соответствие ставился вектор из 300 элементов. Word2Vec – это набор алгоритмов для расчета векторных представлений слов, реализует две основные архитектуры – Continuous Bag of Words (CBOW) и Skip-gram. На вход подается корпус текста, а на выходе получается набор векторов слов. Некоторые применения технологии Word2Vec –определение семантически близких слов, поиск опечаток, оценка важности слов в запросе.

Одно из самых популярных применений нейронных сетей – построение векторов слов, относящихся к области дистрибутивной семантики: считается, что значение слова можно понять по значению его контекста, по окружающим словам. Действительно, если нам незнакомо какое-то слово в тексте на известном языке, то в большинстве случаев можно угадать его значение. Математической моделью значения слова служат вектора слов: строки в большой матрице «слово-контекст», построенной по достаточно большому корпусу текстов. В качестве «контекстов» для конкретного слова могут выступать соседние слова, слова, входящие с данным в одну синтаксическую или семантическую конструкцию, и т. д. В клетках такой матрицы могут быть записаны частоты (сколько раз слово встретилось в данном контексте), но чаще используют коэффициент положительной попарной взаимной информации (Positive Pointwise Mutual Information, PPMI), показывающий, насколько неслучайным было появление слова в том или ином контексте. Такие матрицы вполне успешно могут быть использованы для кластеризации слов или для поиска слов, близких по смыслу к искомому слову.

Как позже было показано, Word2vec – это не что иное, как факторизация матрицы «слово-контекст» с весами PPMI.

3. Лингвистика и обработка текстов

В задачах полнотекстового поиска и классификации текстов не требуется проведения полного морфологического анализа слов, а нужна только проверка того факта, что два указанных слова на самом деле являются формами одного и того же слова. Это может быть выполнено путем их лемматизации (приведение к основной словоформе) или стемминга, который заключается в выделении некоторой неизменяемой части слов. Однако морфологический анализ, лемматизация и стемминг не всегда способны определять родственные слова, например, «безопасность» и «защита». Задачу определения родственных слов решают путем использования специальных словарей-тезаурусов, представляющих собой ориентированные графы, у которых вершины соответствуют словам, а дуги – семантически окрашенным связям между словами. Близость двух слов определяется на основе кратчайшего пути, соединяющего две соответствующие вершины графа. Если необходимо учитывать контекст слов, то задача значительно усложняется, и ее следует отнести к семантической обработке текста. Существуют автоматизированные способы определения связанности слов на основе частоты их совместной встречаемости или степени совпадения их контекстов употребления.

При решении практических задач важно разделить их на две большие группы: обработка отдельных документов и обработка их массивов.

Обработка отдельных текстов

Группу задач обработки отдельных документов следует разделить на две подгруппы: корректировка документов и извлечение информации. Первая подразумевает, что на входе и на выходе будет текстовый документ (задачи исправления ошибок, корректировки текста, определения его структуры, реферирования, машинный перевод). Вторая подгруппа включает в себя задачи, связанные с обработкой формально представленного смысла: извлечение фактов, выполнение запросов на естественном языке, организация естественно-языковых интерфейсов, а также генерация корректных текстов.

Реализация всех задач первой подгруппы может быть основана либо на правилах, сформулированных экспертами, либо на закономерностях, полученных в результате применения методов машинного обучения. Использование системы правил потенциально дает более точные и предсказуемые результаты, но подразумевает большие затраты на ее создание. В свою очередь, применение методов машинного обучения не столь трудоемко, но требует большого количества качественных примеров. Скажем, можно найти множество хороших переводов английского текста на другие европейские языки и обратно, но трудно найти примеры переводов для редких языков.

Вторая подгруппа задач обработки отдельных документов включает в себя извлечение информации, выполнение запросов на естественном языке, генерацию текстов и организацию естественно-языковых интерфейсов. Все эти задачи подразумевают «понимание» текстов и поиск упоминаемых фактов. Работа осуществляется на основе системы правил извлечения информации, каждое из которых задает шаблон синтаксической структуры и шаблон формируемого фрагмента формализованного представления информации. При обработке документа просматриваются результаты синтаксического анализа и ищутся фрагменты, структура которых отвечает шаблонам из правил извлечения информации. Далее в соответствии со «сработавшими» правилами часть слов извлекается из текста и преобразуется в формализованную структуру.

Обработка массивов текстовых документов

Сегодня основная работа с массивами текстовых документов сводится к поиску необходимых документов или их фрагментов, что является обобщением задачи информационного поиска. Данный процесс включает в себя как собственно поиск документов, так и смежные с ним задачи определения ключевых слов, дубликатов документов и цитат, классификации и рубрикации документов, построение кратких обзоров нескольких документов и так далее, вплоть до машинного перевода найденных документов с языка оригинала на язык запроса.

Полнотекстовый поиск в Интернете может рассматриваться как работа с предварительно накопленным хранилищем документов. Существуют задачи, которые рассматривают Сеть как постоянный источник документов и выполняют их «потоковую» обработку. Самыми важными задачами такого типа являются спам-фильтры и системы веб-мониторинга [14].

4. Машинное обучение для NLP

Обработка естественного языка включает в себя распознавание и генерацию речи, классификацию, экстракцию знаний из текстов и другие действия, направленные на понимание текстов с целью наполнения баз знаний, формирования ответов на вопросы и ведения диалога.

Значительный прогресс в области технологий обработки естественного языка в немалой обязан машинному обучению. В машинном обучении особое место принадлежит алгоритмам классификации в задачах, связанных с обработкой текстов: фильтрацией спама, сортировкой документов по тематикам, выделению именованных сущностей. Появилась область тематического моделирования, в которой документы считаются порождением некоего вероятностного процесса и состоят из смеси тем. В компьютерной лингвистике определение частей речи стало высокоточным благодаря таким статистическим методам, как скрытые цепи Маркова и модели максимальной энтропии.

Нейронные сети позволяют находить скрытые связи и закономерности в текстах, но эти связи не могут быть представлены в явном виде. Применение нейронных сетей, во-первых, существенным образом повышает качество решения некоторых стандартных задач классификации текстов и последовательностей, во-вторых, снижает трудоемкость при работе непосредственно с текстами, в-третьих, позволяет решать новые задачи (например, создавать чат-боты).

Нейросетевые технологии коренным образом изменили и работу с текстовыми данными. Если раньше каждый элемент текста (буква, слово или предложение) нужно было описывать с помощью множества признаков различной природы (морфологических, синтаксических, семантических и т. д.), то теперь во многих задачах необходимость в сложных описаниях пропадает. Теоретики и практики нейросетевых технологий часто говорят об «обучении представлению» (representation learning) – в сыром тексте, разбитом только на слова и предложения, нейронная сеть способна найти зависимости и закономерности и самостоятельно составить признаковое пространство. К сожалению, в таком пространстве человек ничего не поймет – во время обучения нейронная сеть ставит каждому элементу текста в соответствие один вектор, состоящих из неких чисел, представляющих обнаруженные «глубинные» взаимосвязи. Акцент при работе с текстом смещается от конструирования подмножества признаков и поиска внешних баз знаний к выбору источников данных и разметке текстов для последующего обучения нейронной сети, для которого требуется существенно больше данных по сравнению со стандартными методами.

Нейронные сети используются в целом ряде задач автоматической обработки текстов (рис. 1).

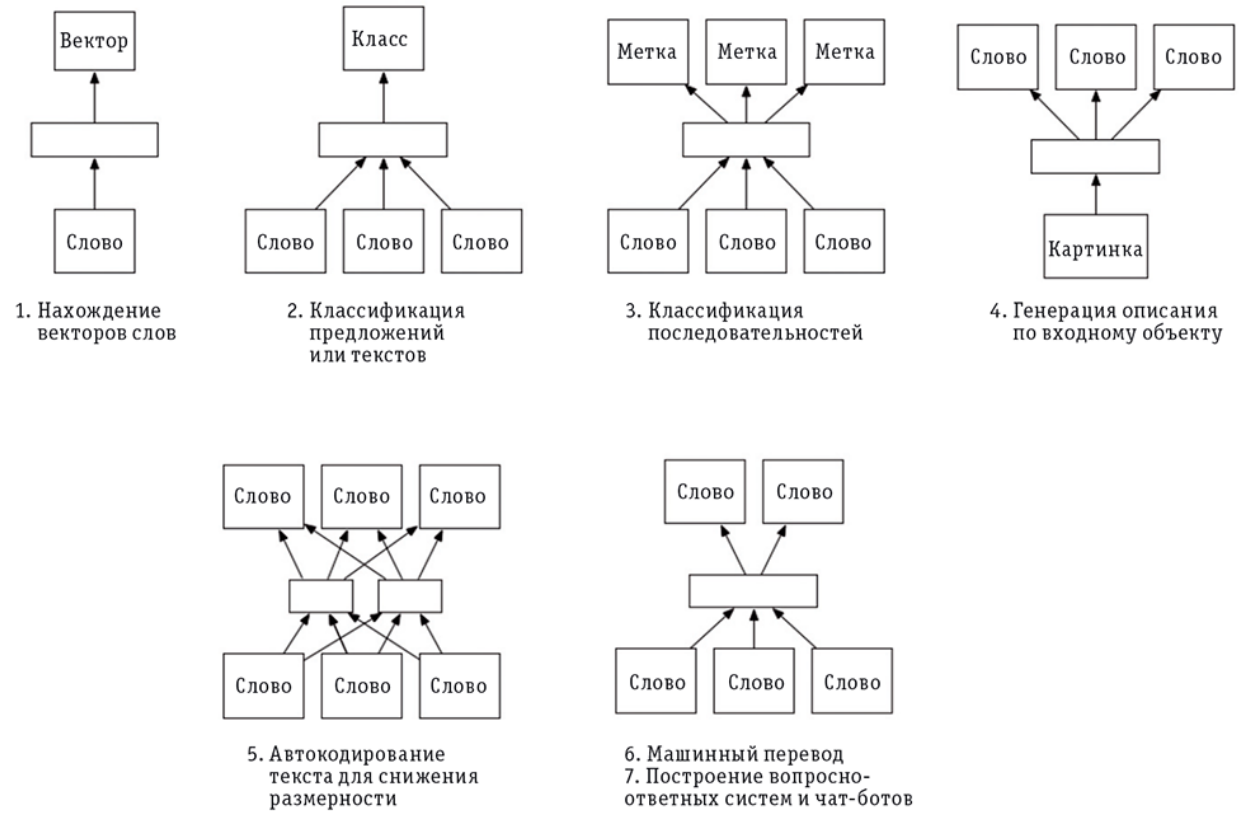

Рис. 1. Архитектуры нейронных сетей, используемые в автоматической обработке текстов [16]

Использование методов глубокого обучения (deep learning), обусловленное прогрессом в области высокопроизводительных систем и появлением больших объемов данных, используемых для обучения, позволило исключить работу по созданию признаков для машинного обучения, предоставляя возможность одновременного обучения выделению признаков и обучения непосредственно самой задаче. Благодаря новым алгоритмам и подходам, включая глубокое обучение, увеличилась скорость грамматического разбора. Кроме того, практически все ведущие алгоритмы и модели стали широко доступны для исследователей. Вероятно, самой известной работой в области глубокого обучения для NLP стал уже упоминавшийся алгоритм Word2vec [12]. Сейчас принято относить Word2vec к дистрибутивной семантике, а не к глубинному обучению, однако исходным толчком для создания этой модели послужило применение нейронной сети. Кроме того, оказалось, что вектора Word2vec служат удобным представлением смысла слова, которое можно подавать на вход глубоким нейронным сетям, используемым для классификации текстов [15].

Представление слов

В традиционном подходе слова предложения обрабатываются как элементы множества слов из словаря, при этом возникают серьезные трудности: если учитывать различные разговорные формы, например, технический жаргон, то объем слов, даже в английском языке, становится слишком большим, а составление полного семантического словаря для широкой области применения – задачей весьма трудоемкой.

Алгоритм Word2vec, входящий во многие стандартные пакеты машинного обучения и обучаемый качественным представлениям слов на больших неразмеченных корпусах (множестве разнообразных текстов по разным темам, написанных в различных жанрах и стилях). В отличие от традиционных представлений слов, здесь используется нейровероятностная модель языка – каждое слово представляется вектором из вещественных чисел в пространстве относительно небольшой размерности (в сравнении с размером полного словаря), например, размерности в 300 измерений. Векторам присваиваются случайные значения, а далее, в процессе обучения, для слова подбирается вектор, максимально похожий (в случае данного алгоритма по мере скалярного произведения) на векторы других слов, которые встречаются в похожих контекстах. В качестве контекста берется небольшое окно предшествующих и

последующих слов (например, пять слов – до, пять – после). Этот подход дает интересные результаты. Во-первых, близкие (по мере скалярного произведения) слова действительно часто семантически близки. Во-вторых, оказывается, что многие интересные для обработки естественного языка отношения закодированы в векторы. Оказалось, что для таких векторов можно определить арифметические операции сложения и вычитания (смыслов). В качестве иллюстрации часто используется следующий пример: если от вектора слова «Париж» отнять вектор слова «Франция» и прибавить вектор «Россия», то получится вектор, очень близкий к вектору «Москва», если вместо вектора «Россия» прибавить вектор «Италия», то результатом будет вектор «Рим» – отношение «столица» оказывается закодированным в векторы слов. Другой пример определяется уравнением: «король – мужчина + женщина = королева».

Как известно, большинство методов обработки естественного языка успешно используют только представления слов, игнорируя синтаксис и семантику, которые можно вывести из синтаксической структуры предложений. Такая модель представления текстов называется BOW (bag of words) – простой набор слов без учета их порядка. Например, в случае векторных представлений можно объединить в кластеры векторы слов корпуса, на которых тренируется модель, и использовать такие кластеры для задач простой классификации. Но если задача состоит в извлечении более качественных семантических представлений, то понадобятся инструменты обработки текстов, работающие с синтаксической структурой предложений или хотя бы не игнорирующие порядок слов в предложении. Например, если требуется анализировать отзывы в социальных медиа о гостиницах, оставленные их клиентами, то можно встретить такое предложение: «Отель приятный, но бар – прокуренный». Без анализа структуры предложения, мы не сможем понять, к какому слову относится каждое прилагательное.

Представление предложений

Методы глубокого обучения предоставляют возможность другого подхода к работе с предложениями – моделирования предложения как последовательности векторов, полученных методом Word2vec, и использования его в алгоритмах машинного обучения.

С такой задачей отлично справляются рекуррентные нейронные сети, которые на входе принимают одно слово в векторном представлении и имеют несколько внутренних уровней, а на выходе строят классификатор.

Классификация последовательностей – это задачи, в которых каждому слову нужно поставить в соответствие одну метку: морфологический разбор (каждому слову ставится в соответствие часть речи), извлечение именованных сущностей (определение того, является ли каждое слово частью имени человека, географического названия и пр.) и т. д. При классификации последовательностей используются методы, позволяющие учитывать контекст слова: если предыдущее слово – часть имени человека, то текущее тоже может быть частью имени, но вряд ли будет частью названия организации. Реализовать это требование на практике помогают рекуррентные нейронные сети, расширяющие идею языковых моделей (language model), предложенных в конце прошлого века. Классическая языковая модель предсказывает вероятность того, что слово i встретится после слова i-1. Языковые модели можно использовать и для предсказания следующего слова: какое слово с наибольшей вероятностью встретится после данного?

Рекуррентные нейросети хорошо себя зарекомендовали при решении различных задач, от моделирования языка до машинного перевода, но у этого класса сетей есть существенный недостаток – они используют только порядок слов в предложении, и их нельзя заставить работать с грамматическими структурами, полученными традиционными инструментами. По сути, рекуррентным сетям приходится для каждой задачи с нуля «учить» грамматику языка.

Кроме того, рекуррентная сеть не строит представлений для промежуточных фраз, поэтому для задач, в которых нужны качественные представления различных фраз, составляющих предложения, используются рекурсивные нейронные сети. При обучении рекурсивная сеть может научиться делать качественные представления не только для полных предложений, но и для всех фраз предложения. При этом нейросеть может ослабить эффект ошибок грамматического разбора, особенно влияющих на задачу, на которой обучается рекурсивная нейросеть. Таким образом, мы получаем меру семантической близости как для слов, так и для всех фраз в предложении.

Развитием рекуррентных нейронных сетей стали архитектуры вида Seq2seq, состоящие из двух соединенных рекуррентных сетей, одна из которых отвечает за представление и анализ входа (например, вопроса или предложения на одном языке), а вторая – за генерацию выхода (ответа или предложения на другом языке). Сети Seq2seq лежат в основе современных систем «вопрос-ответ», чат-ботов и систем машинного перевода.

Другой подход для получения векторов предложений состоит в том, что для каждого предложения, параграфа или целого документа тренируется отдельный вектор, который также участвует в предсказании контекста каждого слова предложения или параграфа, и в процессе обучения выбираются векторы, в наибольшей мере улучшающие предсказания. Название метода – doc2vec. По качеству полученных векторов этот метод соперничает с рекурсивными нейросетями, при этом для обучения не нужна размеченная обучающая выборка. Недостатком метода является требование наличия больших предложений или целых параграфов – он не работает на уровне коротких фраз.

Еще один подход к моделированию слов и предложений основан на использовании сверточных нейросетей. Обычно в сверточных нейросетях на вход сети подается сразу все предложение в виде матрицы векторных представлений отдельных слов. Сверточная сеть обрабатывает произвольно длинную последовательность подсетями фиксированного размера, которые применяются на последовательности окон поверх входных данных. Таким образом эмулируется операция свертки, причем сеть обучается самому фильтру, используемому в свертке. Сверточные сети показывают результаты на уровне рекурсивных сетей.

Кроме сверточных нейронных сетей, для анализа текстов применяются так называемые автокодирующие нейронные сети (autoencoder, автокодировщик), нашедшие применение в лингвистике в задаче снижения размерности (поиск проекции вектора, представляющего текст, на пространство заведомо меньшей размерности). Проекция на двумерное пространство делает возможным представление текста в виде точки на плоскости и позволяет наглядно изобразить коллекцию текстов как множество точек, то есть служит средством предварительного анализа перед кластеризацией или классификацией текстов. Как и сверточные нейронные сети, автокодировщики не имеют никакой лингвистической интерпретации.

С помощью рекуррентных, рекурсивных, сверточных и автокодирующих нейросетей можно эффективно решать задачи, связанные с автоматической обработкой текстов: классификацией, определением тональности, выделением именованных сущностей и простых фактов, снижением размерности и т. д. [15, 16].

5. Практические решения и примеры использования

Сегодня существует множество инструментов для создания систем NLP. Эти инструменты можно условно разбить на три класса: методы работы с индивидуальными словами, методы работы с предложениями и методы для обработки произвольных текстов из нескольких предложений (см. таблицу 1).

Таблица 1. Пакеты для NLP и глубокого обучения [15]

Пакет	Поддерживаемые языки программирования	Статистическая лингвистика	Векторные представления	Глубокое обучение	Источник
Stanford Core NLP	Java	+			stanfornlp.github.io/CoreNLP
NLTK	Python	+			www.nltk.org
OpenNLP	Java	+			opennlp.apache.org
Word2vec	C		+		code.google.com/archive/p/word2vec
Gensim	Python		+		radimrehurek.com/gensim

Таблица 1. Пакеты для NLP и глубокого обучения [15] (Продолжение)

Пакет	Поддерживаемые языки программирования	Статистическая лингвистика	Векторные представления	Глубокое обучение	Источник
Spark Mlib	Java/Scala/Python		+		spark.apache.org/docs/latest/mllib-guide.html
Glove	C		+		nlp.stanford.edu/projects/glove
Google Tensorflow	Python/C++			+	www.tensorflow.org
Theano	Python			+	deeplearning.net/software/theano
Caffe	DSL			+	caffe.berkeleyvision.org

Решение для анализа неструктурированных данных «ПРОМТ»

PROMT Analyser анализирует любые тексты или документы, выделяет в нем сущности (персоналии, организации, географические названия, геополитические сущности и др.), а также определяет соотносящиеся с этими сущностями действия, дату и место совершения действия, формирует целостный образ документа. Система выполняет тонкий морфологический, синтаксический и семантический анализ, что позволяет максимально точно получать информацию из неструктурированных текстовых данных на разных языках, взаимодействуя даже с такими морфологически богатыми, как русский и немецкий, утверждают разработчики. Главным достоинством нового продукта в «ПРОМТ» считают то, что, благодаря инновационным лингвистическим разработкам, он выделяет в текстах также сущности, не представленные в базах. Еще одним преимуществом программы является простая настройка – путем введения значения для новых типов сущностей. Настройка может происходить в том числе и на стороне клиента, что гарантирует ему полную конфиденциальность работы при анализе. Программа, интегрированная в информационно-аналитические и поисковые системы, позволяет решать самые разные задачи в работе с текстовым материалом: анализ внутренних ресурсов (системы документооборота), анализ внешних ресурсов (медиа, блогосфера и пр.), анализ данных, получаемых из закрытых источников для оценки критичности ситуаций, анализ деятельности объекта с привязкой к географии, а также оптимизация поисковых систем и служб поддержки [17].

Семантико-синтаксический анализа текста с помощью технологии Compreno

Одно из направлений деятельности компании ABBYY – потоковый ввод данных в системы корпоративного управления документооборотом, компании принадлежит здесь более половины российского рынка. При этом все чаще корпоративному сектору приходится иметь дело с неструктурированными данными (доля которых по разным оценкам составляет примерно 80%). Для работы с ними в ABBYY предлагают свою технологию «понимания и анализа текстов на естественных языках», получившую название Compreno, которая поможет использовать в бизнесе неструктурированные данные. Новая разработка компании, в отличие от систем, основанных на статистике и правилах, выполняет полный семантико-синтаксический анализ текста, создает его семантическое представление, извлекает из него сущности, события и связи между ними. В частности, технология способна определять по контексту смысл многозначных слов, восстанавливать пропущенные слова, определять связи между объектом, выраженным существительным, и заменяющим его местоимением и т. д.

На базе Compreno в данный момент реализовано три решения в виде комплектов разработчика: Intelligent Search SDK, InfoExtractor SDK и Smart Classifier SDK. В дальнейшем технология может быть оформлена в коробочный продукт [18].

Разработки для автоматического анализа соцсетей: Текстерра и Талисман

Институт системного программирования РАН (ИСП РАН) представил две разработки, предназначенные для анализа больших массивов данных социальных сетей. «Текстерра» – технология многоязычного интеллектуального анализа текста. В ее основе – программный комплекс, способный накапливать и анализировать информацию, устанавливать взаимосвязь между понятиями и проводить интеллектуальный анализ текстов. В результате при использовании этой технологии возможно вычислять лексически сходные тексты, принадлежащие перу одного автора или написанные по единому плану, содержащие сходные семантические конструкции, выявлять факт и структуру информационных «вбросов». Система «Талисман» позволяет отследить и выявить фиктивные аккаунты, служащие для распространения недостоверной информации, спама, имитации массовой активности – ботов. Анализ производится по полным и неполным данным, указанным в профилях социальных сетей, социальным связям аккаунтов, подпискам, спискам друзей и читателей [19].

Категоризация способов использования NLP

В таблице 2 собрано описание способов использования NLP в различных сферах от задач поиска, редактирования текста, анализа его тональности до различных задач юриспруденции и прогноза поведения. В целом, задачи обработки естественного языка встречается повсюду.

Таблица 2. Категоризация способов использования NLP [20]

Поиск	Веб-поиск	Поиск по документам	Автозаполнение
Редактирование	Правописание	Грамматика	Стиль
Диалог	Чат-бот	Помощник	Планирование
Письменный текст	Индексирование	Согласование	Оглавление
Адрес электронной почты	Спам-фильтр	Классификация	Приоритизация
Интеллектуальный анализ текста	Составление краткого содержания	Извлечение знаний	Медицинские диагнозы
Юриспруденция	Правовое влияние	Поиск прецедентов	Классификация повесток в суд
Новости	Обнаружение событий	Проверка фактов	Составление заголовков
Присваивание	Обнаружение плагиата	Литературная экспертиза	Советы по стилю
Анализ тональности текста	Мониторинг морального состояния общества	Сортировка отзывов на продукты	Техническая поддержка
Прогноз поведения	Финансы	Прогноз выборов	Маркетинг
Литературное творчество	Сценарии фильмов	Поэзия	Слова песен

Поисковая машина может предоставить более точные результаты поиска, если индексирует вебстраницы или архивы документов таким образом, что принимается во внимание значение текста на естественном языке. Автозаполнение использует NLP, чтобы закончить вашу мысль. Оно широко распространено в поисковых системах и клавиатурах мобильных устройств. Во множестве текстовых про-

цессоров, плагинов для браузера и текстовых редакторов есть встроенные средства проверки орфографии, грамматики и согласования слов. Некоторые из диалоговых машин (чат-ботов) используют поиск на естественном языке, чтобы найти ответ на сообщение своего собеседника.

Заключение

В настоящее время имеется большое количество актуальных прикладных задач, для решения которых наиболее эффективны методы компьютерной лингвистики и автоматической обработки текстов на естественных языках. В представленном обзоре рассматриваются современные методы и подходы к обработке и анализу текстов, основанные как на применении статистических методов и подходов машинного обучения (включая глубинное обучение), так и на эффективном сочетании различных подходов. Представлены методы семантического анализа текстов, представления слов, предложений и документов, в том числе основанные на представлении в векторном пространстве. Обсуждается разница в подходах и возможностях анализа как отдельных текстов, так и больших массивов текстовой информации. Приведен сравнительный обзор возможностей современных программных инструментов для работы с естественными языками и решений для прикладных направлений (информационно-аналитические и поисковые системы, корпоративный документооборот, автоматический анализ данных из социальных сетей и т.д.). Обобщены типичные задачи для компьютерной лингвистики и анализа текстов в различных областях научной и экономической деятельности.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда (проект № 19-71-30008).

Список литературы

1. Miller, R. AP's 'robot journalists' are writing their own stories now // The Verge, 29 января 2015 года. – URL : www.theverge.com/2015/1/29/7939067/ap-journalism-automation-robots-financial-reporting (дата обращения: 03.02.2020).
2. New York Times, 18 октября 2016 года. – URL : www.nytimes.com/2016/11/18/technology/automatedpro-trumpbots-overwhelmed-pro-clinton-messages-researchers-say.html и MIT Technology Review, ноябрь 2016 года: www.technologyreview.com/s/602817/how-the-bot-y-politicinfluenced-this-election/ (дата обращения: 21.11.2019).
3. Тезаурус РуТез. – URL : http://www.labinform.ru/pub/ruthes/ (дата обращения: 21.11.2019).
4. Word Net: An Electronic Lexical Database / Edit. by Christiane Fellbaum. – Cambridge, MIT Press, 1998 (дата обращения: 5.04.2020).
5. Национальный Корпус Русского Языка. – URL : http://ruscorpora.ru (дата обращения: 21.11.2019).
6. ГИКРЯ: генеральный интернет-корпус русского языка. – URL : http://www.webcorpora.ru/ (дата обращения: 5.04.2020).
7. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова [и др.]. – М. : Изд-во НИУ ВШЭ, 2017. – 269 с.
8. Маннинг, К., Рагхаван, П., Шютце, Ч. Введение в информационный поиск. : Пер. с англ. – М. : ООО «И.Д. Вильямс», 2011. – 528 с.
9. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ [и др.]. – М. : МИЭМ, 2011. – 272 с.
10. Thomas Landauer, Peter W. Foltz, & Darrell Laham. Introduction to Latent Semantic Analysis (англ.) // Discourse Processes (англ.) русск.: journal. – 1998. – Vol. 25. – Pp. 259–284. – DOI: 10.1080/01638539809545028.
11. Mark Needham. scikit-learn: TF/IDF and cosine similarity for computer science papers. – 2017. – URL : https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/ (дата обращения: 21.11.2019).
12. Tomas Mikolov et. al. Efficient Estimation of Word Representations in Vector Space, arxiv.org. – URL: http://arxiv.org/pdf/1301.3781.pdf (дата обращения: 21.11.2019).
13. Stefan Jansen. Word and Phrase Translation with word2vec. – arXiv:1705.03127 [cs.CL]. – 2018. – Pp. 2–11 (дата обращения: 21.11.2019).
14. Селезнев, К. Лингвистика и обработка текстов / К. Селезнев, А. Владимиров // Открытые системы. СУБД. – 2013. – № 4. – С. 46–49.
15. Велихов, П. Машинное обучение для понимания естественного языка // Открытые Системы. СУБД. – 2016. – № 1. – С. 18–21. – URL: https://www.osp.ru/os/2016/01/13048649 (дата обращения: 5.04.2020).
16. Ильвовский, Д. Глубинное обучение для автоматической обработки текстов / Д. Ильвовский, Е. Черняк // Открытые системы. СУБД. – 2017. – № 2. – С. 26–29.
17. «ПРОМТ» представляет решение для анализа неструктурированных данных // Открытые системы. СУБД. – 2016. – № 2. – С. 3–9.
18. Гапотченко, Д. ABBYY учит документооборот понимать тексты // Computerworld Россия. – 2015. – № 10.
19. ИСП РАН анализирует соцсети // Computerworld Россия. – 2017. – № 5.
20. Хобсон, Л. Обработка естественного языка в действии / Л. Хобсон, Х. Ханнес, Х. Коул. – СПб. : Питер, 2020. – 576 с.