Современные технологии обработки естественного языка для решения задач стратегической аналитики*

И. Ф. Кузьминов [I], П. Д. Бахтин [I], А. А. Тимофеев [I], Е. Е. Хабирова [I], П. А. Лобанова [I], Н. И. Зурабьян [II]

[I] Национальный исследовательский университет «Высшая школа экономики», г. Москва, Россия
[II] Государственный научно-исследовательский институт авиационных систем, г. Москва, Россия

Аннотация. Статья посвящена обзору новейших технологий обработки естественного языка (NLP), которые могут быть применены для решения задач стратегической аналитики. Рассмотрены основные проблемы в этой области и конкретные задачи, которые могут быть решены с помощью средств NLP. Приведен обзор основных направлений, в которых задействуются эти средства. Исследованы последние достижения в области NLP и их возможное приложение. Сделаны выводы о том, как должен развиваться аппарат NLP, чтобы в дальнейшем полностью закрыть потребности стратегической аналитики.

Ключевые слова: NLP, искусственный интеллект, текст-майнинг, стратегическая аналитика.

DOI 10.14357/20718594200101

Введение

Стратегическая аналитика – вид научно-практической деятельности, направленный на проведение стратегических исследований и получение ценных для лиц принимающих решения рекомендаций о направлениях политики. Она включает целый ряд направлений, в том числе разработку стратегий социально-экономического, научно-технологического и инновационного развития стран, отраслей и центров компетенций, стратегические маркетинговые исследования (включая исследования рынков, их объемов, структуры и динамики), аналитику по отраслям экономики.

В настоящее время становится все более очевидным нарастающий разрыв между потребностями и возможностями в области стратегической аналитики. Растет потребность систем стратегического управления на государственном и корпоративном уровнях в своевременной и объективной информации (evidence-based policy), суммаризованной до уровня возможности принятия на ее основе конкретных управленческих решений (actionable intelligence). При этом все более явной становится и неспособность традиционных институционально-организационных механизмов, таких как иерархические системы профильных центров компетенций (научных институтов, аналитических центров, консалтинговых компаний), удовлетворить эту потребность. Данную тенденцию по масштабу, всеобщности и негативным последствиям можно охарактеризовать как нарастающий структурный кризис информационного обеспечения системы управления социально-экономической и научно-технологической сферами. Ее причины кроются в беспрецедентном нарастании объемов значи-

* Статья подготовлена в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) с использованием средств субсидии в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

✉ *Бахтин Павел Денисович. E-mail:pbakhin@hse.ru*

мой и релевантной информации на фоне цифровизации глобальной экономики и экономик ведущих стран, включая Российскую Федерацию [1]. Рост объемов ценной и релевантной для принятия решений информации (а не сугубо дублирующихся данных) связан не только с ростом возможности информационных систем фиксировать и отражать большее количество явлений в экономической, технической и иных сферах. Он обусловлен также и структурным усложнением этих сфер, диверсификацией, возникновением большего числа параллельных процессов развития, ведущих параллельную инновационную деятельность коллективов, организаций и иных центров компетенций [2].

Сейчас поиск путей решения изложенной проблемы лежит в области форсайта и исследований будущего. Объединение научных и бизнес-сообществ при активном участии государства в рамках форсайта позволяет приблизиться к созданию единой картины социально-экономического, научного и технологического развития и понять, с какими трудностями и потребностями общество с наибольшей вероятностью столкнется в будущем [3, 4]. Такая стратегия дает возможность частично структурировать разрастающиеся объемы информации путем повышения уровня осведомленности о зарождающихся трендах, в том числе научно-технологических, и выстроить долгосрочное стратегическое планирование в соответствии с этим знанием [5]. К тому же, форсайт, направленный на выявление вызовов и возможностей [6, 7], помогает избегать вызовы, а не искать пути их преодоления в неструктурированном потоке данных [8]. Однако поиск путей решения проблемы нарастания объемов значимой информации с помощью форсайта не дает окончательных ответов.

Традиционные роли в рамках организационных систем экспертной аналитики (такие как руководитель аналитического направления, эксперт предметной области и техник-аналитик) больше не могут эффективно исполняться без опоры на системы автоматического сбора и интеллектуального анализа больших данных [9]. Это связано, в первую очередь, с базовыми биологическими ограничениями естественных способностей человеческого мозга по обработке и запоминанию информации. Применительно к стратегической аналитике, ее автоматизации и аугментации за счет технологий обработки естественного языка, в частности текст-майнинга (первичной обработки больших неструктурированных текстовых массивов) и семантического анализа (смыслового структурирования предобработанных текстовых данных) [10], важным фактором является крайне ограниченная скорость чтения текстов человеком и количество времени, которое человек по базовым физиологическим ограничениям способен заниматься чтением. Также ключевым барьером эффективной аналитики, как адекватного отражения сложности окружающего мира и происходящих в нем процессов, являются особенности работы памяти человека, не позволяющие запоминать в точности достаточно большое количество фактов, связей между ними и их структурно-количественных параметров, отраженных в текстовых документах.

Данное исследование направлено на преодоление проблемы недостаточно своевременного и объективного информирования лиц, принимающих стратегические решения на государственном и корпоративном уровне (в том числе в рамках поддержки формирования стратегий научно-технологического развития стран, отраслей и отдельных системообразующих компаний) об актуальных глобальных, страновых и отраслевых тенденциях развития, прежде всего, научно-технологического. Решение лежит в области автоматизации умственного труда в секторе стратегической аналитики. Наиболее значительный потенциал в этой области связан с обработкой больших документно-текстовых данных и их метаданных (natural language processing, NLP), так как текст, в отличие от других типов больших данных, несет в себе напрямую понимаемый человеком смысл, не требующий сопровождающей легенды (семантического ключа). Именно работа со смысловыми артефактами всегда была и остается ядром любой экспертно-аналитической работы и формирования рекомендаций в области государственной и корпоративной политики и стратегии, включая в первую очередь научно-технологическую сферу как ключевую на сегодняшний день с точки зрения обеспечения конкурентоспособности и возможностей долгосрочного устойчивого развития мира в целом, стран, отраслей и компаний.

В статье представлен обзор текущего состояния сферы NLP с фокусом на возможности этой технологии ответить на потребности стра-

тегической аналитики и решить задачи лиц принимающих решения в госуправлении и корпоративном управлении крупных системообразующих компаний, а также рассмотрены бизнес-приложения и перспективные стратегические приложения NLP.

1. Основные направления развития обработки естественного языка

Компьютерная обработка естественного языка является одним из ключевых направлений искусственного интеллекта по стратегическому трансформационному потенциалу. Амбициозные цели в сфере NLP заявлялись давно (например, в 1950-х гг. достижение совершенного машинного перевода ожидалось к началу 1960-х гг.), но реальная применимость NLP-технологий росла постепенно. Радикальный прорыв пришелся лишь на 2010-е годы, когда количество патентов в области искусственного интеллекта многократно возросло по сравнению с предыдущими периодами [11].

На сегодняшний день объем глобального рынка решений и сервисов NLP оценивается экспертами в более 10 млрд долл. с прогнозом более 22 млрд долл. в 2025 г. [12] Наиболее востребованными применениями NLP-технологий являются распознавание речи, информационный поиск (эволюционирующий в направлении рекомендательных систем), синтез текста, машинный перевод, а также извлечение ценных сведений из больших данных (текст-майнинг), в т.ч. для стратегической аналитики, принятия высокоуровневых управленческих решений в государственном и корпоративном секторе.

Распознавание речи – самостоятельное крупное направление искусственного интеллекта (с оценкой рынка более 12 млрд долл. в 2019 г. с прогнозом более 24 млрд долл. в 2025 г. [13]). В нем NLP-технологии используются для решения части задач, улучшения результатов, но в ряде случаев они не требуются. Например, возможно распознавание речи для отдельных задач (голосовое распознавание команд в интерфейсах пилотирования самолетов нового поколения и др.) за счет использования больших обучающих выборок аудиосэмплов с опорой только на последовательности звуков, анализ частотного спектра и амплитуды звуковых колебаний, без использования сложных моделей NLP, отражающих структуру речи на уровне слов, частей речи, синтаксических связей, предложений и т.д. Однако с учетом роста популярности голосового управления при взаимодействии с гаджетами и бытовыми приборами, а также конкуренции на рынках, побуждающей производителей техники делать свою продукцию все более комфортной в использовании [14], спрос на NLP-технологии в этой сфере будет только расти.

Рекомендательные системы также являются отдельным направлением искусственного интеллекта (с оценкой рынка более 1,5 млрд долл. в 2019 г. с прогнозом более 12,5 млрд долл. в 2025 г. [15]) и только частично опираются на технологии NLP. Остальные названные направления меньше по объему рынка, но теснее интегрированы с NLP (в особенности синтез текста, машинный перевод).

Системы, решающие все названные выше задачи, интегрируют широкий спектр технологий NLP, обеспечивающих базовый пайплайн обработки текста. К ним относятся ETL, классификация и кластеризация источников, сегментация текста, токенизация, морфологический анализ, лемматизация/стеммизация, синтаксический разбор и выявление именованных сущностей, семантический анализ, статистический анализ данных всех предыдущих этапов обработки, сентимент-анализ, связывание сущностей, суммаризация, предсказание и генерация текста. Эти интегративные системы используются далее в рамках глобальных цепочек создания стоимости как техническое ядро (платформенные B2B сервисы) для многих конечных коммерческих приложений B2B, B2G и B2C. К их числу относятся, в порядке убывания оцененного, в том числе экспертно, перспективного объема, следующие рынки:

– NLP для здравоохранения и наук о жизни, прежде всего для оптимизации поисковых исследований в фармакологии, биомедицине, например, для автоматизации поиска лекарственных кандидатов, сокращения пространства перебора комбинаций молекул в органической химии за счет текст-майнинга научных статей, отчетов по клиническим испытаниям; «майнинг генов»; управление медицинскими знаниями, предсказательная медицина на основе интеллектуального анализа lifelong историй болезни и др. Оценка рынка: более 1,5 млрд долл. в 2019 г. с прогнозом более 5,5 млрд долл. в 2025 г. [16].

– Conversational AI: чат-боты, или естественно-языковые интерфейсы, а также платформы для конструирования кастомизированных чат-ботов (Dialogflow, Rasa NLU/Core и др.). Представлены широким спектром решений от простых, основанных на правилах ботов, предназначенных для проведения пользователя по строго заданному бизнес-процессу, до сложных решений, обеспечивающих поддержание естественной беседы на широкий спектр тематик и при этом отслеживающих ключевые моменты беседы, где пользователь упоминает точку входа в конкретный бизнес-процесс. В будущем, наряду с графическим пользовательским интерфейсом и интерфейсом командной строки, чат-боты станут третьим, неотъемлемым видом интерфейса для большинства информационных систем. Однако их применения, особенно перспективные, этим далеко не ограничиваются. Так, показана возможность поддержания осмысленного, информативного диалога между двумя чат-ботами, в том числе за счет использования новых эмбеддинговых моделей, таких как BERT. Этот функционал имеет не только чисто академический интерес. Напротив, в будущем он может применяться для таких прикладных задач, как моделирование фокус-групп, экспертных дискуссионных панелей, форсайт-сессий, использующих цифровые модели реальных экспертов, а также для обучения риторическим навыкам в рамках образовательных процессов, в спецтренингах подготовки к бизнес-переговорам и др. Оценка рынка чат-ботов: более 0,5 млрд долл. в 2019 г. с прогнозом более 3,5 млрд долл. в 2025 г. [17].

– Маркетинг 3.0 и другие приложения распознавания эмоций, прежде всего системы, обеспечивающие вовлечение потребителя, значимый обмен идеями с обширной аудиторией продукта в рамках коротких, двухнедельных agile-циклов. Важно обеспечение эффективной интерактивной коммуникации за счет майнинга намерений, агрегации и интеллектуального анализа отзывов в реальном времени, сентимент-анализа продуктов и услуг и их отдельных аспектов (aspect-based sentiment analysis) по данным социальных сетей. Оценка рынка: более 0,2 млрд долл. в 2019 г. с прогнозом более 3,5 млрд долл. в 2025 г. (самые быстрые прогнозируемые темпы роста среди рынков NLP [18]).

– Разнообразные конечные пользовательские приложения машинного перевода, в т.ч. перспективные сервисы синхронного перевода устной речи (последние особенно актуальны, в числе прочего, для современных многопользовательских компьютерных игр, предотвращая излишнюю сегментацию игрового мира и повышая тем самым потребительскую ценность, расширяя рынок за счет дополнительных категорий потребителей). Оценка рынка машинного перевода: более 0,5 млрд долл. в 2019 г. с прогнозом более 1 млрд долл. в 2025 г. [19].

– HR Tech: автоматизированное принятие кадровых решений, умное соотнесение резюме с вакансиями, оценка кандидата по его текстовому следу, выявление неочевидных склонностей и талантов, гибкое формирование проектных команд и др. Оценка рынка HR Tech: более 0,5 млрд долл. в 2019 г. с прогнозом более 1 млрд долл. в 2025 г., однако только часть этого рынка связана непосредственно с технологиями NLP [20].

– Новый ритейл: магазины без продавцов, виртуальные помощники в интернет-магазинах, интеграция взаимодействия с клиентами и логистики поставок в единую автоматизированную цепочку, реагирующую на события и адаптирующуюся к среде в режиме реального времени для оптимального решения задачи одновременной максимизации удовлетворенности потребителей, доли рынка, продаж и прибылей.

– Эволюция поисковиков веба (Google, DuckDuckGo, Bing, Yandex) в рекомендательные системы (наподобие Wolfram Alpha, но с лучшей реализацией и с сохранением функции поиска веб-сайтов). Трансформация привычных поколению Y-функций и бизнес-моделей поисковиков веба происходит достаточно давно. Она шла в направлении от сугубо технического информационного поиска на основе выдачи формально релевантных ключевым словам ссылок (инвертированный индекс, PageRank, tf-idf, LSI и т.п.) к выдаче пользователю результатов, которые были ранее успешными для подобных ему пользователей с похожими поисковыми запросами. Затем эволюция повернула в направлении отражения компонентов графа знаний, соответствующих истинному намерению пользователя (на основе intention mining), и предоставления сопутствующих информационных сервисов. Пользователю становится не нужно переходить на другие веб-сайты, он получает нужную ему сводку (Google OneBox – первый шаг в этом направлении, получивший

развитие с внедрением Сети знаний или Knowledge Graph [21]), удовлетворяет все свои потребности, в том числе в общении и покупках, непосредственно на сайте веб-поисковика. Вероятна массовая негативная реакция владельцев индексируемых веб-сайтов на дальнейшее развитие этого тренда. Ответом может стать создание ИТ-гигантами собственных «фабрик контента» и новой модели рекламной монетизации в сети, от которой выиграют ИТ-гиганты и потребители и проиграют независимые рекламные агентства и владельцы традиционных веб-сайтов.

– Автоматизация поисковой оптимизации, «семантическое SEO», «этическое SEO».

– Legal Tech: поисково-рекомендательные системы, автоматизирующие ряд задач, традиционно выполнявшихся младшими юристами, таких как поиск релевантных кейсов в странах с прецедентным правом, извлечение и структуризация ценных данных из документов (например, извлечение ковенантов контрактов на основе правил, форм-ридинга, семантики с их автоматической конвертацией в структурированные форматы данных).

– Fraud detection, антифрод, due diligence, автоматические проверки заемщиков, финансовый контроль (например, синтаксический разбор и семантический анализ коротких текстов о назначении платежа в платежных поручениях на предмет выявления паттернов, характерных для мошенничества и т.п.).

– Ed Tech: индивидуальные обучающие системы, обеспечивающие персонализованную виртуальную среду обучения и умную обратную связь для студента, а также системы поддержки деятельности преподавателей.

– Робожурналистика: автоматический синтез и стилизация текста на основе постоянно пополняемых хранилищ данных о происходящих событиях; рынок шаблонов для генерации новостей разных типов по разным отраслям; распространение тренда робожурналистики за пределы финансовых и спортивных новостей на другие сферы, более сложные с точки зрения автоматизации нарративов. Идея создания устройства, позволявшего генерировать тексты автоматически, высказывалась еще Джонатаном Свифтом в «Приключениях Гулливера» [22]. Перспективы аппарата NLP в этой области косвенно подтверждаются результатами применения более примитивных моделей SCIgen [23], Mathgen [24] и Paper generator [25], создающих текст по принципу контекстно-свободной грамматики.

– Интегрированные аналитические среды для проведения расследований, обработки разведочных данных, обеспечивающие адаптивную загрузку и умное объединение разнородных, в том числе текстовых материалов (Palantir, IBM i2, SAS Text Analytics). Сюда же стоит отнести проект Google Forest Watch, предоставляющий информацию о росте и вырубке лесов на основании разнородных данных [26].

– Более умные блокировщики рекламы, в том числе так называемые этичные эдблокеры, продвинутые спам - и скам-фильтры.

– Суммаризация: извлечение основной информации из длинных текстов, ее обобщение и представление в удобном, привычном для читателя формате.

– Облегчение работы исследователей по широкому спектру научных направлений (за пределами фармакологии и медицины, на которые приходится львиная доля рынка NLP для научных исследований) за счет рекомендаций материалов для чтения, структурирования данных научных статей, умных сводок и т.д.

– Повышение качества распознавания рукописного ввода, OCR.

– Частичная автоматизация формирования онтологий для узких предметных областей.

– Развитие межъязыковых переводов для редких языков.

– Генерация свободных нарративов (например, частичная, в будущем вероятно полная автоматизация формирования деревьев диалогов с персонажами в компьютерных играх с большим миром; автоматизация производства кратких описаний продуктов, в том числе книг, фильмов; генерация вариантов питчей для репетиции публичных выступлений и т.п.).

– Process mining, анализ и мониторинг функционирования сложных информационных систем по данным больших массивов логов (применение NLP ограничено, так как даже когда данные текстовые, как правило, они строго структурированы, с предельно простой ограниченной грамматикой, однако это направление можно рассматривать как предтечу зарождающегося поднаправления NLP – ALP, т.е. обработки *искусственного* языка – см. ниже).

– Обработка искусственных языков, отличающихся более простой и предсказуемой

структурой, большей семантической однозначностью, чем естественный язык (например, языки программирования, разметки данных и др.). Речь идет в том числе о нейронных сетях, генерирующих программный код, а также о перспективных мультиязычных семантических моделях языков программирования для облегчения автоматической трансляции кода. Универсальная трансляция кода, машинное понимание и генерация инструкций на искусственных языках на основе технологий ALP позволит решить многие проблемы неэффективности, асимметрии в ИТ-индустрии, которые не могут в полной мере решить парное программирование [27] и другие существующие техники.

– Многочисленные малые, нишевые направления NLP, возникающие либо перспективные, пока не существующие, зачастую для малого и среднего бизнеса. Удовлетворение таких потребностей в NLP будет наиболее эффективным сначала через уникальные сервисы стартапов, а затем по мере роста зрелости услуги – через недорогие массовые подписки на стандартизованные облачные NLP-решения в пакете прочих облачных сервисов (в рамках таких платформ, как MS Azure, Amazon AWS, Digital Ocean и подобных).

2. Ключевые технологические тренды в сфере NLP

Драйвером удовлетворения спроса потребителей на перечисленных выше рынках является постоянный научно-технический прогресс сферы NLP, рост технологических возможностей анализа данных, точности и многогранности результатов предсказательных моделей, скорости и надежности их обучения и работы. В последние годы наблюдается технологическая революция в NLP [28] благодаря развитию глубинного обучения [29, 30] и появлению эмбеддинговых моделей, присваивающих каждому устойчивому буквосочетанию, слову, термину языка и даже документу [31] числовой вектор большой размерности, позволяя тем самым предсказывать элемент по его контексту или контекст по элементу, рассчитывать смысловую близость понятий и документов [32], производить «семантические арифметические операции» [33-35]. Параллельно наблюдается стремительная оптимизация таких моделей (минимальная модель английского языка в фреймворке spaCy для синтаксического и морфологического анализа занимает всего 17МБ, почти не уступая по эффективности моделям прошлых поколений, имевшим размер до нескольких гигабайт).

Одновременно появляются новые поколения многослойных эмбеддинговых моделей, пока очень большие по объему и вычислительно сложные, но радикально расширяющие возможности NLP [36]. Речь идет о моделях и подходах BERT [37], ULMfit [28], GPT-2 [38], ELMo [28], позволяющих кодировать смысл слов и добавляющих к возможностям ставших традиционными эмбеддинговых моделей (таких как word2vec, GloVe, fastText) новые, генеративные возможности (заполнение пропусков в тексте, написание расширенного текста по начальному фрагменту [39], ответы на вопросы с обоснованием выбора). В качестве промежуточного звена между этими поколениями можно принять модель lda2vec [40], еще не позволявшую полноценно кодировать смысл слов, но представлявшую их в виде векторов, координатами которых являлись вероятности принадлежности к той или иной теме. Исследования границ возможностей этих моделей продолжаются по сей день, а также разработана для сентимент-анализа модель ABAE [41], где слова характеризуются принадлежностью к тому или иному аспекту. Ярким примером является недавняя успешная попытка генерации с помощью GPT-2 текстов патентных заявок [42]. Возникают межъязыковые эмбеддинговые модели, открывающие путь к машинному переводу без опоры на обучение по дорогостоящим параллельным текстовым корпусам.

Векторные представления активно вытесняют трудоемкие алгоритмы (в частности, требующие feature engineering) в традиционных для NLP задачах (морфологический, синтаксический анализ, распознание именованных сущностей, сентимент-анализ, классификация, slot filling). Эмбеддинги наряду с нейронными сетями разных архитектур (рекуррентные, конволюционные, с долгой краткосрочной памятью, перестраиваемые и др.) и разными способами применения нейросетей [43] стали в последние годы новой технологической базой NLP на логическом уровне. Модели на их основе легко визуализируются, например, с помощью инструментария DxR [44]. Для их реализации ис-

пользуется программное и аппаратное обеспечение более низкого уровня. В частности, созданная на базе сиамских нейросетей модификация BERT – Sentence-BERT (SBERT) позволила существенно сократить время поиска пары предложений с наибольшим семантическим сходством по сравнению с классическим BERT. Модель предполагает сравнение векторных представлений целых предложений с использованием косинус-подобия, в результате чего время поиска наиболее схожей пары в наборе из 11000 предложений сокращается с ~65 часов до ~5 секунд [45].

Универсальным языком разработки (зачастую инструментом «склейки», единым интерфейсом) для NLP-решений становится Python (с низкоуровневыми алгоритмами, чаще всего реализованными на C++ или Cython). Для этого языка реализованы алгоритмы кластеризации OPTICS [46] и HDBSCAN [47]. Задачи хранения текстовых данных и результатов их обработки все лучше решаются noSQL-решениями (Elasticsearch [48], Dgraph [21] и др.), появляется возможность отказа от сложных в развертывании и поддержке решений с распределенными реляционными базами данных (такие, как шардинг в Postgres). Внедряются архитектуры, предусматривающие контейнеризацию (Docker, Kubernetes) и растет популярность старых, но эффективных форматов для абстрагирования внутренней логики приложений и сервисов (REST API, JSON как универсальный формат обмена данными). Архитектура NLP-решений будет развиваться в сторону модульности. Предоставление NLP-технологий пользователям будет мигрировать в сторону модели SaaS, XaaS. Все это обусловливает стремительное моральное и техническое устаревание более ранних NLP-решений, написанных на теряющих популярность языках, на закрытых или не поддерживаемых открытым сообществом библиотеках, использующих менее эффективные решения по машинному обучению (метод опорных векторов вместо более современных нейросетевых технологий в синтаксическом анализе и др.). Такие устаревающие, но еще занимающие значительную долю рынка приложения нередко опираются на традиционные нераспределенные клиент-серверные архитектуры, самописные базы данных с реляционной логикой и соответственно высокой вычислительной стоимостью многоуровневых запросов, плохой масштабируемостью, дороговизной поддержки, сложностью управления изменениями и, как следствие, неизбежным поступательным замедлением развития [49]. В будущем выживут в коммерческом плане только проекты, реализующие лучшие модели и концепции организации рабочего процесса в ИТ: Lean IT, Agile и DevOps. Добьются успеха NLP-проекты, открытые к постоянной миграции на новые архитектуры, фреймворки и даже языки ради постоянного роста гибкости, масштабируемости, эффективности коммуникации на всех уровнях, в том числе с сообществом свободного программного обеспечения.

Из конкретных примеров практического применения новейших технологий в сфере NLP можно выделить разработанную в Национальном университете оборонных технологий Китая модель суммаризации текста на основе BERT. Ее эффективность подтверждена исследованием, где в качестве входных данных использовались текстовые корпуса CNN/Daily Mail и New-York Times [50]. Другие примеры использования подхода BERT для суммаризации текстов описаны в работах [51, 52].

Модели и подходы, позволяющие генерировать текст автоматически, стали толчком к созданию моделей, позволяющих отличать сгенерированный текст от написанного человеком [53]. Ярким примером является модель Grover [54].

Продолжается и эволюция самих эмбеддинговых моделей нового поколения. В июне 2019 г. появилась модель XLNet, превзошедшая BERT в решении ряда задач [55], а уже в июле этого же года – модель RoBERTa [56], созданная специалистами Facebook на базе усовершенствованного подхода BERT и превосходящая по возможностям XLNet. Информация о новейших разработках в этом направлении публикуется на ресурсах Papers With Code [57] и NLP-progress [58].

3. Применение NLP в Российской Федерации

В России разработки в области NLP затрагивают многие сферы. В частности, проводилось исследование, которое показало, что с его помощью можно автоматически выявлять тексты противоправного содержания и экстремистской направленности на сайтах и в соцсетях [59]. В 2018 г. было проведено исследование, показавшее, что по текстам и профилям в соцсетях

можно эффективно выявлять у людей депрессивное состояние [60].

Одним из ярких примеров применения NLP в сфере веб-технологий является сценарный чат-бот, созданный компанией «Наносемантика»[1]. Общение с ним по характеру гораздо ближе к разговору с живым онлайн-консультантом, чем в случае кнопочного чат-бота. Наиболее широкий спектр решений на базе NLP в этой области предлагают такие компании, как NAUMEN[2] и ООО «ЭР СИ О»[3].

В последнее время активно исследуются возможности применения NLP и искусственного интеллекта в сфере медицины [61, 62]

В сфере маркетинга заслуживает внимания платформа PolyAnalyst, с помощью которой можно проводить анализ отзывов о компаниях для получения детальной оценки по различным аспектам и сравнения с конкурентами [63].

Ярким примером автоматизации умственного труда в сфере стратегической аналитики является система Exactus Expert, позволяющая анализировать большие объемы научных документов [64]. Кроме этого, в 2016г. был запущен проект Ростелекома «Мониторинг трендов», в основе которого лежит анализ большого количества текстовой информации, включая патенты и научные публикации. Благодаря тому, что в основе применяемой методики лежит машинное обучение, роль экспертов в процессе мониторинга трендов удалось существенно сократить [65]. К этой же сфере относится большая часть решений, предлагаемых «Айкумен ИБС», основателем которой стал один из ведущих специалистов в России в области искусственного интеллекта Сергей Шумский. В 2015 г. сделка по продаже «Айкумен» Ростелекому стала первой в России покупкой NLP-проекта [66]. Решения, предлагаемые «Айкумен ИБС», созданы на базе поисково-аналитической платформы IQPLATFORM[4]. Кроме того, на базе NLP были разработаны технологии автоматического поиска заимствований в научных текстах [67]. В этом же направлении активные разработки ведутся на Факультете компьютерных Наук НИУ ВШЭ [68]. Одной из последних значимых разработок стала модель анализа структуры научных сообществ [69].

[1] https://nanosemantics.ai
[2] https://www.naumen.ru/
[3] http://www.rco.ru/
[4] https://www.iqmen.ru/

Заключение

Развитие направления NLP и приложений автоматического обучения машин (AutoML) в NLP ведет к существенному снижению зависимости от массового экспертного труда (онтологический инжиниринг и разметка данных могут быть очень дороги, до 2 млрд долл., как, например, было в случае с группами стандартов ISO-15926 и ISO-25964, зародившимися из амбициозного онтологического проекта в нефтегазовой отрасли). Снижается и зависимость от дорогого труда специалистов по интеллектуальному анализу данных (ручной подбор параметров, feature engineering). Согласно прогнозам, со временем человеческий фактор будет практически полностью вытеснен из многих сфер, включая и перевод текстов любой тематики [66].

Однако прогресс пока еще не дошел до такого уровня, чтобы отпала необходимость в ресурсах ручной разметки данных, экспертного формирования онтологий. Тем временем, даже эмбеддинговые модели и нейронные сети в задачах NLP, пока не позволяют автоматически генерировать точные предметные таксономии и онтологии, не дают, особенно для экспертных задач, тех результатов, которые возможны на основе дорогостоящих онтологически-контролируемых методов (условно говоря, пока невозможны «автоматически генерируемые экспертные системы on-demand»). Соответственно, встает главный технологический вопрос современного NLP: что придет после эмбеддингов и как это поможет в решении по-настоящему востребованных и пока принципиально нерешаемых прикладных задач для искусственного интеллекта в этой сфере? Ответа на этот вопрос пока нет.

Однако ясны сами конечные потребности, которые пока еще не могут быть удовлетворены с помощью ИИ [70], но которые в будущем, вероятно, смогут быть решены за счет более продвинутых технологий NLP, объединенных с другими ИИ-технологиями для комплексной обработки разных типов больших данных: текстовых, графических, аудио, данных Интернета вещей, статистических материалов из разрозненных официальных источников, с финансовых бирж, цифровых моделей оборудования и др.

Так, в первую очередь, остается недоступной имитация креативности на основе машин-

ного понимания данных вообще [71] (в том числе текстовых – NLU), которая была бы полезна в решении практических управленческих, экспертных и инженерных задач. Примером тому может служить недавняя попытка применить ИИ к произведениям классической литературы [72]. Недостижим пока надежный синтез новых практически полезных идей из больших данных, частичное замещение инженерных коллективов в поисковых исследованиях, оценке технических возможностей, проектировании нового работоспособного и эффективного оборудования по заданным спецификациям (первые маленькие шаги в этом направлении предприняты в рамках парадигмы generative design). Тем более, невозможно пока автоматическое формирование самих требований, совмещающих амбициозность и реалистичность, к новым поколениям техники, S.M.A.R.T.-спецификаций на основе комплексного осмысления больших данных, описывающих достигнутые возможности современной науки и техники, доступные материалы, процессы, алгоритмы.

Еще более сложными задачами для ИИ, использующего, в том числе NLP, является содействие в принятии высокоуровневых стратегических решений не просто на уровне предоставления информации к размышлению, а на уровне формирования конкретных предложений по политике, стратегии, приоритетам, управленческим шагам в социально-экономической сфере, с обоснованием на естественном языке предлагаемых инициатив.

Названные задачи, какими бы фантастическими они ни казались с точки зрения сегодняшнего уровня ИИ вообще и NLP в частности, отражают реальные потребности органов власти и крупных компаний, высокопоставленных лиц принимающих решения ведущих экспертов и лидеров ключевых инженерных коллективов. Технологические решения для ответа на эти потребности должны быть найдены, иначе повторится ситуация, которая сложилась к 1973 г. [73] и очередной «зимы ИИ» будет не избежать. Поэтому ИИ-компании, желающие добиться успеха в долгосрочной перспективе, должны формулировать новые стратегические постановки в сфере развития NLP не от достигнутого, а от глубинных долгосрочных трансформационных потребностей государства, бизнеса, общества и человека.

Литература

1. Went P. How does the digital economy create ‘alternative data’? [Электронный ресурс]. URL: https://link.medium.com/dCTNLPPuEW (дата обращения 02.10.2019 г.).
2. Кузьминов И. Ф., Логинова И. В., Лобанова П. А. Перспективы использования технологий анализа больших данных для стратегической аналитики агропромышленного комплекса // Сахарная свекла. 2018. № 9. С. 2-7.
3. Соколов А. В., Чулок А. А. Долгосрочный прогноз научно-технологического развития России на период до 2030 года: ключевые особенности и первые результаты // Форсайт. 2012. Т. 6. № 1. С. 12-25.
4. Osipov G. et al. Information retrieval for R&D support //Professional search in the modern world. Springer. Cham, 2014. P. 45-69.
5. Meissner D. Approaches for Developing National STI Strategies // STI Policy Review. 2014. Vol. 5. No. 1. P. 34-56.
6. King D. A., Thomas S. M. Taking science out of the box--foresight recast //Science. 2007. Т. 316. №5832. P. 1701-1702.
7. Martin B. R. Foresight in science and technology //Technology analysis & strategic management. 1995. Т. 7. №2. P. 139-168.
8. Sokolov A. et al. Future of S&T: Delphi survey results //Foresight and STI Governance (Foresight-Russia till No. 3/2015). 2009. Т. 3. №3. P. 40-58.
9. Кузьминов И. Ф., Лобанова П. А., Логинова И. В. Технология анализа больших данных для стратегической аналитики отрасли // Комбикорма. 2019. № 4. P. 46-52.
10. Berry M. W. Survey of text mining / M. W. Berry // Computing Reviews. 2004. Т. 45. № 9. P. 548.
11. Patents in Artificial Intelligence - 1969-2017. [Электронный ресурс]. URL: https://www.econsight.ch/wp-content/ai/index.html (дата обращения 02.10.2019 г.).
12. MarketsandMarkets. Natural Language Processing Market worth 16.07 Billion USD by 2021. [Электронный ресурс]. URL: https://www.marketsandmarkets.com/PressReleases/natural-language-processing-nlp.asp (дата обращения 02.10.2019 г.).
13. Grand View Research. Voice and Speech Recognition Market Size, Share & Trends Analysis Report, By Function, By Technology (AI, Non-AI), By Vertical (Healthcare, BFSI, Automotive), And Segment Forecasts, 2018 – 2025. [Электронный ресурс]. URL: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market (дата обращения 02.10.2019 г.).
14. Kim M. 2019 UI and UX Design Trends. [Электронный ресурс]. URL: https://uxplanet.org/2019-ui-and-ux-design-trends-92dfa8323225 (дата обращения 14.10.2019).
15. MarketsandMarkets. Recommendation Engine Market by Type (Collaborative Filtering, Content-Based Filtering, and Hybrid Recommendation), Deployment Mode (Cloud and On-Premises), Technology, Application, End-User, and Region - Global Forecast to 2022. [Электронный ресурс]. URL: https://www.marketsandmarkets.com/Market-Reports/recommendation-engine-market-151385035.html (дата обращения 02.10.2019 г.).

16. MarketsandMarkets. Natural Language Processing (NLP) in Healthcare and Life Sciences Market by Component (Technology and Services), Type (Rule-based, Statistical and Hybrid), Application, Deployment Mode (Cloud and On Premise) and Region - Global Forecast to 2021. [Электронный ресурс]. URL: https://www.marketsandmarkets.com/Market-Reports/healthcare-lifesciences-nlp-market-131821021.html (дата обращения 02.10.2019 г.).
17. Mordor Intelligence. Chatbots Market Size - Segmented by Type (Solution, Service), Deployment (On-Premise, Cloud), End-User Vertical (BFSI, Healthcare, IT and Telecommunication, Retail, Utilities, Government), and Region - Growth, Trends and Forecast (2019 - 2024). [Электронный ресурс]. URL: https://www.mordorintelligence.com/industry-reports/chatbots-market (дата обращения 02.10.2019 г.).
18. Tractica. Emotion Recognition and Sentiment Analysis Market to Reach $3.8 Billion by 2025. [Электронный ресурс]. URL: https://www.tractica.com/newsroom/press-releases/emotion-recognition-and-sentiment-analysis-market-to-reach-3-8-billion-by-2025/ (дата обращения 02.10.2019 г.).
19. Grand View Research. Machine Translation Market Size To Reach $983.3 Million by 2022. [Электронный ресурс]. URL: https://www.grandviewresearch.com/press-release/global-machine-translation-market (дата обращения 02.10.2019 г.).
20. HR Technologist. HR Tech Marketplace Worth $400 Billion. [Электронный ресурс]. URL: https://www.hrtechnologist.com/news/digital-transformation/hr-tech-marketplace-worth-400-billion/ (дата обращения 14.10.2019).
21. Jain M. R. Why Google Needed a Graph Serving System. [Электронный ресурс]. URL: https://blog.dgraph.io/post/why-google-needed-graph-serving-system/ (дата обращения 14.10.2019)..
22. Garcia C. Gulliver’s engine. [Электронный ресурс]. URL: https://computerhistory.org/blog/gullivers-engine/?key=gullivers-engine (дата обращения 14.10.2019).
23. SCIgen - An Automatic CS Paper Generator. [Электронный ресурс]. URL: https://pdos.csail.mit.edu/archive/scigen/ (дата обращения 14.10.2019).
24. Mathgen. [Электронный ресурс]. URL: http://thatsmathematics.com/mathgen/ (дата обращения 14.10.2019).
25. Laplante P.A. 3.7.5 Paper Generators // Technical Writing: A Practical Guide for Engineers and Scientists. CRC Press. 2011. P. 56–59.
26. Gibson J. From 0 to 2 million in 4 years. [Электронный ресурс]. URL: https://medium.com/vizzuality-blog/global-forest-watch-from-0-to-2-million-in-4-years-32f63cd9a46 (дата обращения 14.10.2019).
27. Leung J. The Benefits of Pair Programming. [Электронный ресурс]. URL: https://medium.com/better-programming/when-pair-programming-works-it-works-really-well-heres-why-c51857bbcf0f (дата обращения 14.10.2019).
28. Saravia E. NLP 2018 Highlights. [Электронный ресурс]. URL: http://elvissaravia.com/nlp-highlights-2018/ (дата обращения 14.10.2019).
29. Transformers from scratch. [Электронный ресурс]. URL: http://www.peterbloem.nl/blog/transformers (дата обращения 14.10.2019).
30. Ильвовский Д., Черняк Е. Глубинное обучение для автоматической обработки текстов // Открытые системы. СУБД. 2017. № 2. С. 26-29.
31. Lau J. H., Baldwin T. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation // Proceedings of the 1st Workshop on Representation Learning for NLP. 2016. P.78-86.
32. STSbenchmark. [Электронный ресурс]. URL: http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark (дата обращения 14.10.2019).
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality, in Proceedings of NIPS [Электронный ресурс]. URL: https://arxiv.org/abs/1310.4546 (дата обращения 14.10.2019).
34. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in VectorSpace. – CoRR, abs/1301.3781, 2013 – URL: http://arxiv.org/abs/1301.3781.
35. Huber F. King - Man + Woman = King? [Электронный ресурс]. URL: https://blog.esciencecenter.nl/king-man-woman-king-9a7fd2935a85 (дата обращения 14.10.2019).
36. Singh C. Fine-Tune ERNIE 2.0 for Text Classification. [Электронный ресурс]. URL: https://towardsdatascience.com/https-medium-com-gaganmanku96-fine-tune-ernie-2-0-for-text-classification-6f32bee9bf3c (дата обращения 14.10.2019).
37. Rajasekharan A. Deconstructing BERT. [Электронный ресурс]. URL: https://towardsdatascience.com/deconstructing-bert-reveals-clues-to-its-state-of-art-performance-in-nlp-tasks-76a7e828c0f1 (дата обращения 14.10.2019).
38. How To Make Custom AI-Generated Text With GPT-2. [Электронный ресурс]. URL: https://minimaxir.com/2019/09/howto-gpt2/ (дата обращения 14.10.2019).
39. Write With Transformer. [Электронный ресурс]. URL: https://transformer.huggingface.co/ (дата обращения 14.10.2019).
40. Ma E. Combing LDA and Word Embeddings for topic modeling. [Электронный ресурс]. URL: https://towardsdatascience.com/combing-lda-and-word-embeddings-for-topic-modeling-fe4a1315a5b4 (дата обращения 14.10.2019).
41. He R., Lee W.S., Ng H.T., Dahlmeier D. An Unsupervised Neural Attention Model for Aspect Extraction // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. P.388-397.
42. Lee J.-S., Hsiang J. Patent Claim Generationby Fine-Tuning OpenAI GPT-2. [Электронный ресурс]. URL: https://arxiv.org/abs/1907.02052 (дата обращения 14.10.2019).
43. Richter M. Comparing Word Embeddings. [Электронный ресурс]. URL: https://towardsdatascience.com/comparing-word-embeddings-c2efd2455fe3 (дата обращения 14.10.2019).
44. Hackathorn R. DxR: Bridging 2D Data Visualization into Immersive Spaces. [Электронный ресурс]. URL:

https://towardsdatascience.com/dxr-bridging-2d-data-visualization-into-immersive-spaces-d77a20d5f9e9 (дата обращения 14.10.2019).
45. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [Электронный ресурс]. URL: https://arxiv.org/pdf/1908.10084.pdf (дата обращения 14.10.2019).
46. Sinclair C. Clustering Using OPTICS. [Электронный ресурс]. URL: https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7 (дата обращения 14.10.2019).
47. Oskolkov N. How to cluster in High Dimensions. [Электронный ресурс]. URL: https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6 (дата обращения 14.10.2019).
48. Kopichinsky G. Improve Heavy Elasticsearch Aggregations with Random Score and Sampler Aggregation. [Электронный ресурс]. URL: https://medium.com/cognigo/improve-heavy-elasticsearch-aggregations-with-random-score-and-sampler-aggregation-9e1857271059 (дата обращения 14.10.2019).
49. Gutteridge L. What I'm Telling Business People About Why Relational Databases Are So Bad. [Электронный ресурс]. URL: https://codeburst.io/what-im-telling-business-people-about-why-relational-databases-are-so-bad-6f38d3d6c995 (дата обращения 14.10.2019).
50. Liu Y., Lapata M. Text Summarization with Pretrained Encoders. [Электронный ресурс]. URL: https://www.researchgate.net/publication/335337738_Text_Summarization_with_Pretrained_Encoders (дата обращения 14.10.2019).
51. Liu Y. Fine-tune BERT for Extractive Summarization. [Электронный ресурс]. URL: https://www.researchgate.net/publication/331986865_Fine-tune_BERT_for_Extractive_Summarization (дата обращения 14.10.2019).
52. Zhang H., Xu J., Wang J. Pretraining-Based Natural Language Generation for Text Summarization. [Электронный ресурс]. URL: https://arxiv.org/pdf/1902.09243.pdf (дата обращения 14.10.2019).
53. Metz C., Blumental S. How A.I. Could Be Weaponized to Spread Disinformation. [Электронный ресурс]. URL: https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html (дата обращения 14.10.2019).
54. Grover — A State-of-the-Art Defense against Neural Fake News. [Электронный ресурс]. URL: https://grover.allenai.org/ (дата обращения 14.10.2019).
55. Saravia E. XLNet outperforms BERT on several NLP Tasks. [Электронный ресурс]. URL: https://medium.com/dair-ai/xlnet-outperforms-bert-on-several-nlp-tasks-9ec867bb563b (дата обращения 14.10.2019).
56. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. [Электронный ресурс]. URL: http://arxiv.org/abs/1907.11692 (дата обращения 14.10.2019).
57. Trending Research. [Электронный ресурс]. URL: https://paperswithcode.com/ (дата обращения 14.10.2019).
58. NLP-progress. [Электронный ресурс]. URL: http://nlpprogress.com/ (дата обращения 14.10.2019).
59. Ананьева М.И., Девяткин Д.А., Кобозева М.В., Смирнов И.В., Соловьев Ф.Н., Чеповский А.М. Исследование характеристик текстов противоправного содержания // Труды ИСА РАН. 2017. Т. 67. № 3. С.86-97.
60. Станкевич М.А., Исаков В.А., Девяткин Д.А., Смирнов И.В. Построение классификационных моделей для задачи обнаружения депрессии у пользователей социальных сетей // В сборнике: Информатика, управление и системный анализ Труды V Всероссийской научной конференции молодых ученых с международным участием. 2018. С.237-246.
61. Сомс Н.Л., Добров А.В. AI-технологии NLU и Ontological Semantics в медицинских экспертных системах. [Электронный ресурс]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/3.pdf (дата обращения 14.10.2019).
62. Ефименко И.В. Семантический анализ текстов в области медицины и биотеха: проблемы и перспективы. [Электронный ресурс]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/6.pdf (дата обращения 14.10.2019)
63. Русских А.Н. Анализ отзывов клиентов о ведущих лабораторных провайдерах. [Электронный ресурс]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/4.pdf (дата обращения 14.10.2019).
64. Osipov G. et al. Exactus expert—search and analytical engine for research and development support // Novel Applications of Intelligent Systems. – Springer, Cham, 2016. – С. 269-285.
65. Мониторинг трендов. [Электронный ресурс]. URL: https://digitaltrends.rt.ru (дата обращения 14.10.2019).
66. Под ред. Шумского С.А. – Пивоваров И.О. и др. Альманах «Искусственный интеллект». [Электронный ресурс]. URL: http://www.aireport.ru/ (дата обращения 14.10.2019).
67. Осипов Г.С., Смирнов И.В., Тихомиров И.А., Соченков И.В., Зубарев Д.В., Исаков В.А. Технологии семантического поиска заимствований в научных текстах // Книга. Культура. Образование. Инновации («Крым-2016») Материалы Второго Международного профессионального форума. 2016. С.311-313.
68. Ена О.В., Нагаев К.В. Автоматизация процессов разработки технологических дорожных карт. Расчет интегральных показателей применимости // Бизнес-информатика. 2013. № 3 (25). С.56-62.
69. Кузьминов И. Ф., Бахтин П. Д., Незнанов А. А., Лобанова П. А. Исследования структуры научного сообщества на основе семантического анализа: выявление и кластеризация центров компетенций и тематик // В кн.: Управление научными исследованиями и разработками. Государство и наука: новые модели управления – 2018. – Труды Четвертой научно-практической конференции (26 ноября 2018 г., Москва). ИПУ РАН. 2019. С.128-137.
70. McClory P. Is AI Our Last Hope for a Big Disruption? Or Just The Newest One? [Электронный ресурс]. URL: https://towardsdatascience.com/is-ai-our-last-hope-for-a-big-disruption-or-just-the-newest-one-357f9c3db618 (дата обращения 14.10.2019).
71. Giacaglia G. The Road to Artificial General Intelligence. [Электронный ресурс]. URL: https://medium.com/datadriveninvestor/the-road-to-artificial-general-intelligence-cfcb37bdc432 (дата обращения 14.10.2019).

72. Kolakowski N. Unleashing Machine Learning on Literature's Great Works. [Электронный ресурс]. URL: https://link.medium.com/kfFBclDJHW (дата обращения 14.10.2019).
73. Schuchmann S. History of the first AI Winter. [Электронный ресурс]. URL: https://link.medium.com/SSfaFF1PGW (дата обращения 14.10.2019).

Modern Natural Language Processing Technologies for Solving Strategic Analytics Tasks

I. F. Kuzminov[I], P. D. Bakhtin[I], A. A. Timofeev[I], E. E. Khabirova[I], P. A. Lobanova[I], N. I. Zurabyan[II]

[I] National Research University Higher School of Economics, Moscow, Russia
[II] State Research Institute of Aviation Systems, Moscow, Russia

Abstract. The article is devoted to a review of the latest natural language processing (NLP) technologies that can be applied in strategic analytics. The introduction discusses the main problems in this area and specific tasks that can be solved using NLP tools. The article provides an overview of the main application areas in which these tools are involved. The paper reviews recent advancements in NLP and assess their potential. Conclusions are drawn about how the NLP apparatus should be developed in order to fulfill the needs of strategic analytics in the future.
Keywords: NLP, artificial intelligence, text mining, strategic analytics.

DOI 10.14357/20718594200101

References

1. Went P. How does the digital economy create 'alternative data'? [Electronic resource]. URL: https://link.medium.com/dCTNLPPuEW (accessed 02.10.2019).
2. Kuz'minov I. F., Loginova I. V., Lobanova P. A. Perspektivy ispol'zovanija tehnologij analiza bol'shih dannyh dlja strategicheskoj analitiki agropromyshlennogo kompleksa // Saharnaja svekla, 2018. – № 9. – P. 2-7.
3. Sokolov A. V., Chulok A. A. Dolgosrochnyj prognoz nauchno-tehnologicheskogo razvitija Rossii na period do 2030 goda: kljuchevye osobennosti i pervye rezul'taty // Forsajt. 2012. – T. 6. – № 1. – P. 12-25.
4. Osipov G. et al. Information retrieval for R&D support //Professional search in the modern world. – Springer, Cham, 2014. – P. 45-69.
5. Meissner D. Approaches for Developing National STI Strategies // STI Policy Review. 2014. – Vol. 5. – No. 1. – P. 34-56.
6. King D. A., Thomas S. M. Taking science out of the box--foresight recast //Science. – 2007. – T. 316. – №. 5832. – P. 1701-1702.
7. Martin B. R. Foresight in science and technology //Technology analysis & strategic management. – 1995. – T. 7. – №. 2. – P. 139-168.
8. Sokolov A. et al. Future of S&T: Delphi survey results //Foresight and STI Governance (Foresight-Russia till No. 3/2015). – 2009. – T. 3. – №. 3. – P. 40-58.
9. Kuz'minov I. F., Lobanova P. A., Loginova I. V. Tehnologija analiza bol'shih dannyh dlja strategicheskoj analitiki otrasli // Kombikorma. 2019. – № 4. – P. 46-52.
10. Berry, M. W. Survey of text mining / M. W. Berry // Computing Reviews. — 2004. — T. 45. — № 9. — P. 548.
11. Patents in Artificial Intelligence - 1969-2017. [Electronic resource]. URL: https://www.econsight.ch/wp-content/ai/index.html (accessed 02.10.2019).
12. MarketsandMarkets. Natural Language Processing Market worth 16.07 Billion USD by 2021. [Electronic resource]. URL: https://www.marketsandmarkets.com/PressReleases/natural-language-processing-nlp.asp (accessed 02.10.2019).
13. Grand View Research. Voice and Speech Recognition Market Size, Share & Trends Analysis Report, By Function, By Technology (AI, Non-AI), By Vertical (Healthcare, BFSI, Automotive), And Segment Forecasts, 2018 – 2025. [Electronic resource]. URL: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market (accessed 02.10.2019).
14. Kim M. 2019 UI and UX Design Trends. [Electronic resource]. URL: https://uxplanet.org/2019-ui-and-ux-design-trends-92dfa8323225 (accessed 14.10.2019).
15. MarketsandMarkets. Recommendation Engine Market by Type (Collaborative Filtering, Content-Based Filtering, and Hybrid Recommendation), Deployment Mode (Cloud and On-Premises), Technology, Application, End-User, and Region - Global Forecast to 2022. [Electronic resource]. URL: https://www.marketsandmarkets.com/Market-Reports/recommendation-engine-market-151385035.html (accessed 02.10.2019).
16. MarketsandMarkets. Natural Language Processing (NLP) in Healthcare and Life Sciences Market by Component (Technology and Services), Type (Rule-based, Statistical

and Hybrid), Application, Deployment Mode (Cloud and On Premise) and Region - Global Forecast to 2021. [Electronic resource]. URL: https://www.marketsandmarkets.com/Market-Reports/healthcare-lifesciences-nlp-market-131821021.html (accessed 02.10.2019).
17. Mordor Intelligence. Chatbots Market Size - Segmented by Type (Solution, Service), Deployment (On-Premise, Cloud), End-User Vertical (BFSI, Healthcare, IT and Telecommunication, Retail, Utilities, Government), and Region - Growth, Trends and Forecast (2019 - 2024). [Electronic resource]. URL: https://www.mordorintelligence.com/industry-reports/chatbots-market (accessed 02.10.2019).
18. Tractica. Emotion Recognition and Sentiment Analysis Market to Reach $3.8 Billion by 2025. [Electronic resource]. URL: https://www.tractica.com/newsroom/press-releases/emotion-recognition-and-sentiment-analysis-market-to-reach-3-8-billion-by-2025/ (accessed 02.10.2019).
19. Grand View Research. Machine Translation Market Size To Reach $983.3 Million by 2022. [Electronic resource]. URL: https://www.grandviewresearch.com/press-release/global-machine-translation-market (accessed 02.10.2019).
20. HR Technologist. HR Tech Marketplace Worth $400 Billion. [Electronic resource]. URL: https://www.hrtechnologist.com/news/digital-transformation/hr-tech-marketplace-worth-400-billion/ (accessed 14.10.2019).
21. Jain M. R. Why Google Needed a Graph Serving System. [Electronic resource]. URL: https://blog.dgraph.io/post/why-google-needed-graph-serving-system/ (accessed 14.10.2019).
22. Garcia C. Gulliver's engine. [Electronic resource]. URL: https://computerhistory.org/blog/gullivers-engine/?key=gullivers-engine (accessed 14.10.2019).
23. SCIgen - An Automatic CS Paper Generator. [Electronic resource]. URL: https://pdos.csail.mit.edu/archive/scigen/ (accessed 14.10.2019).
24. Mathgen. [Electronic resource]. URL: http://thatsmathematics.com/mathgen/ (accessed 14.10.2019)..
25. Laplante P.A. 3.7.5 Paper Generators // Technical Writing: A Practical Guide for Engineers and Scientists. – CRC Press, 2011. – P. 56–59.
26. Gibson J. From 0 to 2 million in 4 years. [Electronic resource]. URL: https://medium.com/vizzuality-blog/global-forest-watch-from-0-to-2-million-in-4-years-32f63cd9a46 (accessed 14.10.2019).
27. Leung J. The Benefits of Pair Programming. [Electronic resource]. URL: https://medium.com/better-programming/when-pair-programming-works-it-works-really-well-heres-why-c51857bbcf0f (accessed 14.10.2019).
28. Saravia E. NLP 2018 Highlights. [Electronic resource]. URL: http://elvissaravia.com/nlp-highlights-2018/ (accessed 14.10.2019).
29. Transformers from scratch. [Electronic resource]. URL: http://www.peterbloem.nl/blog/transformers (accessed 14.10.2019).
30. Il'vovskij D., Chernjak E. Glubinnoe obuchenie dlja avtomaticheskoj obrabotki tekstov // Otkrytye sistemy. – SUBD. 2017. – № 2. – P. 26-29.
31. Lau J. H., Baldwin T. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation // Proceedings of the 1st Workshop on Representation Learning for NLP. – 2016. – S. 78-86.
32. STSbenchmark. [Electronic resource]. URL: http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark (accessed 14.10.2019).
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality, in Proceedings of NIPS [Electronic resource]. URL: https://arxiv.org/abs/1310.4546 (accessed 14.10.2019).
34. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in VectorSpace. – CoRR, abs/1301.3781, 2013 – URL: http://arxiv.org/abs/1301.3781.
35. Huber F. King - Man + Woman = King? [Electronic resource]. URL: https://blog.esciencecenter.nl/king-man-woman-king-9a7fd2935a85 (accessed 14.10.2019).
36. Singh C. Fine-Tune ERNIE 2.0 for Text Classification. [Electronic resource]. URL: https://towardsdatascience.com/https-medium-com-gaganmanku96-fine-tune-ernie-2-0-for-text-classification-6f32bee9bf3c (accessed 14.10.2019).
37. Rajasekharan A. Deconstructing BERT. [Electronic resource]. URL: https://towardsdatascience.com/deconstructing-bert-reveals-clues-to-its-state-of-art-performance-in-nlp-tasks-76a7e828c0f1 (accessed 14.10.2019).
38. How To Make Custom AI-Generated Text With GPT-2. [Electronic resource]. URL: https://minimaxir.com/2019/09/howto-gpt2/ (accessed 14.10.2019).
39. Write With Transformer. [Electronic resource]. URL: https://transformer.huggingface.co/ (accessed 14.10.2019).
40. Ma E. Combing LDA and Word Embeddings for topic modeling. [Electronic resource]. URL: https://towardsdatascience.com/combing-lda-and-word-embeddings-for-topic-modeling-fe4a1315a5b4 (accessed 14.10.2019).
41. He R., Lee W.S., Ng H.T., Dahlmeier D. An Unsupervised Neural Attention Model for Aspect Extraction // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2017. – P.388-397.
42. Lee J.-S., Hsiang J. Patent Claim Generationby Fine-Tuning OpenAI GPT-2. [Electronic resource]. URL: https://arxiv.org/abs/1907.02052 (accessed 14.10.2019).
43. Richter M. Comparing Word Embeddings. [Electronic resource]. URL: https://towardsdatascience.com/comparing-word-embeddings-c2efd2455fe3 (accessed 14.10.2019).
44. Hackathorn R. DxR: Bridging 2D Data Visualization into Immersive Spaces. [Electronic resource]. URL: https://towardsdatascience.com/dxr-bridging-2d-data-visualization-into-immersive-spaces-d77a20d5f9e9 (accessed 14.10.2019).
45. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. [Electronic

resource]. URL: https://arxiv.org/pdf/1908.10084.pdf (accessed 14.10.2019).
46. Sinclair C. Clustering Using OPTICS. [Electronic resource]. URL: https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7 (accessed 14.10.2019).
47. Oskolkov N. How to cluster in High Dimensions. [Electronic resource]. URL: https://towardsdatascience.com/how-to-cluster-in-high-dimensions-4ef693bacc6 (accessed 14.10.2019).
48. Kopichinsky G. Improve Heavy Elasticsearch Aggregations with Random Score and Sampler Aggregation. [Electronic resource]. URL: https://medium.com/cognigo/improve-heavy-elasticsearch-aggregations-with-random-score-and-sampler-aggregation-9e1857271059 (accessed 14.10.2019).
49. Gutteridge L. What I'm Telling Business People About Why Relational Databases Are So Bad. [Electronic resource]. URL: https://codeburst.io/what-im-telling-business-people-about-why-relational-databases-are-so-bad-6f38d3d6c995 (accessed 14.10.2019).
50. Liu Y., Lapata M. Text Summarization with Pretrained Encoders. [Electronic resource]. URL: https://www.researchgate.net/publication/335337738_Text_Summarization_with_Pretrained_Encoders (accessed 14.10.2019).
51. Liu Y. Fine-tune BERT for Extractive Summarization. [Electronic resource]. URL: https://www.researchgate.net/publication/331986865_Fine-tune_BERT_for_Extractive_Summarization (accessed 14.10.2019).
52. Zhang H., Xu J., Wang J. Pretraining-Based Natural Language Generation for Text Summarization. [Electronic resource]. URL: https://arxiv.org/pdf/1902.09243.pdf (accessed 14.10.2019).
53. Metz C., Blumental S. How A.I. Could Be Weaponized to Spread Disinformation. [Electronic resource]. URL: https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html (accessed 14.10.2019).
54. Grover — A State-of-the-Art Defense against Neural Fake News. [Electronic resource]. URL: https://grover.allenai.org/ (accessed 14.10.2019).
55. Saravia E. XLNet outperforms BERT on several NLP Tasks. [Electronic resource]. URL: https://medium.com/dair-ai/xlnet-outperforms-bert-on-several-nlp-tasks-9ec867bb563b (accessed 14.10.2019).
56. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. [Electronic resource]. URL: http://arxiv.org/abs/1907.11692 (accessed 14.10.2019).
57. Trending Research. [Electronic resource]. URL: https://paperswithcode.com/ (accessed 14.10.2019).
58. NLP-progress. [Electronic resource]. URL: http://nlpprogress.com/ (accessed 14.10.2019).
59. Anan'eva M.I., Devjatkin D.A., Kobozeva M.V., Smirnov I.V., Solov'ev F.N., Chepovskij A.M. Issledovanie harakteristik tekstov protivopravnogo soderzhanija // Trudy Instituta sistemnogo analiza Rossijskoj akademii nauk. – 2017. – T. 67. – № 3. – P. 86-97.
60. Stankevich M.A., Isakov V.A., Devjatkin D.A., Smirnov I.V. Postroenie klassifikacionnyh modelej dlja zadachi obnaruzhenija depressii u pol'zovatelej social'nyh setej // V sbornike: Informatika, upravlenie i sistemnyj analiz Trudy V Vserossijskoj nauchnoj konferencii molodyh uchenyh s mezhdunarodnym uchastiem. – 2018. – P. 237-246.
61. Soms N.L., Dobrov A.V. AI-tehnologii NLU i Ontological Semantics v medicinskih jekspertnyh sistemah. [Electronic resource]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/3.pdf (accessed 14.10.2019).
62. Efimenko I.V. Semanticheskij analiz tekstov v oblasti mediciny i bioteha: problemy i perspektivy. [Electronic resource]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/6.pdf (accessed 14.10.2019).
63. Russkih A.N. Analiz otzyvov klientov o vedushhih laboratornyh provajderah. [Electronic resource]. URL: https://armit.ru/medsoft/2019/presentation/Day_01/16_18/4.pdf (accessed 14.10.2019).
64. Osipov G. et al. Exactus expert—search and analytical engine for research and development support // Novel Applications of Intelligent Systems. – Springer, Cham, 2016. – P. 269-285.
65. Monitoring trendov. [Electronic resource]. URL: https://digitaltrends.rt.ru (accessed 14.10.2019).
66. Pod red. Shumskogo S.A. – Pivovarov I.O. i dr. Al'manah «Iskusstvennyj intellekt». [Electronic resource]. URL: http://www.aireport.ru/ (accessed 14.10.2019).
67. Osipov G.S., Smirnov I.V., Tihomirov I.A., Sochenkov I.V., Zubarev D.V., Isakov V.A. Tehnologii semanticheskogo poiska zaimstvovanij v nauchnyh tekstah // Kniga. Kul'tura. Obrazovanie. Innovacii ("Krym-2016") Materialy Vtorogo Mezhdunarodnogo professional'nogo foruma. – 2016. – P. 311-313.
68. Ena O.V., Nagaev K.V. Avtomatizacija processov razrabotki tehnologicheskih dorozhnyh kart. Raschet integral'nyh pokazatelej primenimosti // Biznes-informatika. – 2013. – № 3 (25). – P. 56-62.
69. Kuz'minov I. F., Bahtin P. D., Neznanov A. A., Lobanova P. A. Issledovanija struktury nauchnogo soobshhestva na osnove semanticheskogo analiza: vyjavlenie i klasterizacija centrov kompetencij i tematik // V kn.: Upravlenie nauchnymi issledovanijami i razrabotkami. Gosudarstvo i nauka: novye modeli upravlenija – 2018. – Trudy Chetvertoj nauchno-prakticheskoj konferencii (26 nojabrja 2018 g., Moskva). – IPU RAN, 2019. – P. 128-137.
70. McClory P. Is AI Our Last Hope for a Big Disruption? Or Just The Newest One? [Electronic resource]. URL: https://towardsdatascience.com/is-ai-our-last-hope-for-a-big-disruption-or-just-the-newest-one-357f9c3db618 (accessed 14.10.2019).
71. Giacaglia G. The Road to Artificial General Intelligence. [Electronic resource]. URL: https://medium.com/datadriveninvestor/the-road-to-artificial-general-intelligence-cfcb37bdc432 (accessed 14.10.2019).
72. Kolakowski N. Unleashing Machine Learning on Literature's Great Works. [Electronic resource]. URL: https://link.medium.com/kfFBclDJHW (accessed 14.10.2019)
73. Schuchmann S. History of the first AI Winter. [Electronic resource]. URL: https://link.medium.com/SSfaFF1PGW (accessed 14.10.2019).