DOI: 10.17805/trudy.2024.5.3 *ИНФОРМАТИКА*

# РАБОТА С ТЕКСТОМ: ПРОШЛОЕ, НАСТОЯЩЕЕ, БУДУЩЕЕ

*О. В. Гаврилова*
*Московский гуманитарный университет*

*П. А. Зубковский, IPSOS*

***Аннотация:*** *В статье рассматривается история развития технологий создания, обработки текста, а также — перспективы развития этого процесса.*
***Ключевые слова:*** *текст; текстовый процессор; искусственный интеллект*

# WORKING WITH TEXT: PAST, PRESENT, FUTURE

*O. V. Gavrilova*
*Moscow University for the Humanities*

*P. A. Zubkovsky, IPSOS*

***Abstract:*** *The article examines the history of the development of technologies for creating and processing text, as well as the prospects for the development of this process.*
***Keywords:*** *text; word processor; artificial intelligence*

Современные средства цифровых телекоммуникаций позволяют эффективно воздействовать на сознание населения, а управляющие средства массовой информации, становясь более доступными в реализации цифрового контента, манипулируют поведением населения в требуемых заказчиком направлениях и действенных проявлениях (Костина, 2022). Однако, все началось с Гутенберга. Хотя техника печати для воспроизведения текста, рисунков и изображений была известна в Древнем Китае в III веке нашей эры, это была так называемая штучная печать. К XV в. в Европе возросла потребность книг, но рукописные книги были дороги. И Иоганн Генсфляйш цур Ладен цум Гутенберг в 1439 г. стал первым европейцем, который использовал подвижные литеры для печати. На созданном им станке можно было изготовить 100 оттисков одного листа за час.

Первопечатных книг сохранилось мало. И они очень похожи на рукописные. И, возможно, их и продавали иногда по цене рукописных. Подражание рукописям имело цель не только получить хорошую прибыль. Общество подозревало в печати вмешательство дьявола, и приходилось скрывать способ производства книги. Поэтому на первых печатных экземплярах не было так называемых выходных данных.

Европа не признавала первенство немцев в изобретении книгопечатания. Например, голландцы настаивали, что первым был их соотечественник Лауренс Янсзон Костер. Но на процесс книгопечатания это никак не повлияло. И книги пошли большим потоком в массы.

С ростом грамотности населения у некоторой его части возникло непреодолимое желание видеть свои мысли на печатном листе. И вот в 1714 г. некий английский водопроводчик Генри Милль запатентовал не только прибор, но и сам способ последовательной печати символов. Патент на изобретение «Машины для расшифровки письма» был получен. Применялся ли этот патент или нет — история умалчивает. А затем в самом начале XIX в. произошла романтическая история. Итальянский механик и изобретатель Пеллегрино Турри любил графиню Каролину Фантони, которая внезапно ослепла. И Турри в 1801 г. придумал для нее печатающее устройство, позволяющее вести переписку. Также Турри придумал копировальную бумагу.

Наша страна тоже внесла вклад в развитие этого направления. В России в 1870 г. М. И. Алисов создал «Скоропечатник», наборно-пишущую машину, которая демонстрировалась на выставке в Вене. Примерно тогда же бразильский священник Франсиско Жуан де Азеведо изобрел свой вариант печатной машинки, за что бразильское правительство вручило ему золотую медаль (Шомракова, Баренбаум, 2008).

Сначала пишущие машинки воспринимались как экзотика. Они были дороги и малочисленны (как рукописные книги). Тут подключились американцы. Журналист, изобретатель и печатник Кристофер Лэтем Шоулз и механик любитель Карлос С. Глидден создали в 1873 г. первую серийно выпускавшуюся и коммерчески успешную печатную машинку. В 1877 г. Кристофер Шоулз продал права на изготовление своей усовершенствованной печатной машинки компании Remington.

Среди первых покупателей «Ремингтона» были писатели Марк Твен и Лев Толстой. Считается, что первым литературным произведением, напечатанным на пишущей машинке, был «Том Сойер». Пишущих машинок стало очень много. Много стало и машинисток.

Иногда в различных СМИ возникает тревожная мысль о том, что же будет с представителями разных профессий при автоматизации производственных процессов. Не грозит ли прогресс массовой безработицей? Например, по дороге от Москвы до Санкт-Петербурга уже поехали грузовики на автопилоте. А куда же делось большое количество машинисток, которые еще недавно трудились в каждой организации?

Технология механического печатания имела много недостатков. Ввод и правка текста требовали много времени. Копии создавались только с помощью копировальной бумаги. Шрифт для данной пишущей машинки был фиксированный. И человечество с появлением компьютеров придумало текстовые редакторы, а затем и текстовые процессоры.

Отличие текстовых процессоров от текстовых редакторов заключается в следующем: при работе с текстовыми файлами текстовый редактор осуществляет просмотр их содержимого и производит вставку, удаление и копирование текста, сортировку строк, печать. А текстовый процессор включает в текст иллюстрации, формирует различного рода указатели и ссылки, вводит колонтитулы

страниц, производит поиск текста и исправление ошибок, копирует и переносит в другой документ фрагменты текста.

Первый текстовый редактор был создан в 1964 г. и назывался QED (Quick Editor). Разработан он был Барри Вейном (Barry Weymiller) для использования на компьютере CDC 1604. Майкл Шрейер, 20 лет занимавшийся телерекламой и киносъёмками, купил в 1975 г. MITS Altair 8800 и начал программировать в качестве хобби, не рассматривая это как источник заработка. В 1980-х гг. текстовые редакторы стали более популярными, благодаря появлению персональных компьютеров и операционной системы Microsoft Windows. Microsoft Word, выпущенный в 1983 г., стал наиболее распространенным.

Важным свойством программного обеспечения, работающего с текстом, является реализация концепции WYSIWYG. Концепция WYSIWYG предполагает графический интерфейс пользователя, позволяющий ему заранее видеть объект творчества еще на этапе создания. Аббревиатура WYSIWYG означает: what you see is what you get — «что ты видишь, то и получишь», характеризует программное обеспечение, позволяющее редактировать визуальное представление информации в форме близкой к его конечной реализации, например, печатного документа, веб-страницы или слайд-презентации. В контексте работы с текстовыми документами WYSIWYG подразумевает возможность прямого формирования макета документа без необходимости использовать специальные команды для работы с макетом.

До реализации концепции WYSIWYG для отображения текста в редакторах использовался основной системный шрифт и стиль с упрощенным описанием макета — полей, интервалов и многое другое. При редактировании документа следовало добавлять в поле с текстом непечатаемые управляющие коды — тегами кода разметки, определяющих детали конечного отображения. В каждой программе был свой собственный язык разметки, требующий больших затрат времени на освоение пользователем.

Использование тегов и кодов разметки по-прежнему популярно в некоторых приложениях, благодаря их способности хранить сложную информацию о форматировании. Однако освоение таких редакторов требует больших трудозатрат, к тому же усложняется контроль пользователя над конечным видом документа.

Первым текстовым редактором с использованием подхода WYSIWYG стал Bravo для компьютера Alto, созданный в научно-исследовательском центре Xerox PARC в 1974 г., отображающая текст с форматированием (например, с выравниванием по ширине, шрифтами и пропорциональным интервалом между символами). На мониторе компьютера Alto можно было увидеть целую страницу текста. Текст отображался на экране с разрешением 72 точки на дюйм, в то время как при печати использовалось разрешение 300 точек на дюйм. В результате некоторые символы немного искажались. Несмотря на то, что проблема легкого несоответствия отображения документа на экране и при печати пусть в меньшей мере, но присутствует и сегодня, концепция WYSIWYG лишь набирает свою популярность.

Хорошим примером альтернативного, пусть и устаревшего, но еще популярного подхода является TeX — программа набора текста, разработанная специалистом по информатике и профессором Стэнфордского университета Дональдом Кнутом. В основе TeX лежит язык разметки, определяющий форматирование текста. Первый релиз TeX появился в 1978 г.

Сегодня TeX — популярное средство для набора сложных математических формул, однако не может составить конкуренцию редакторам и программам верстки, использующих принцип WYSIWYG. TeX популярен в научных кругах, особенно среди физиков, математиков и других специалистов, публикующих работы, изобилующих сложными формулами и требующих соответствующей верстки.

При обработке текста важна коррекция ошибок. Алгоритмы этого процесса постоянно усложняются (Основы информационных технологий, 2010). В 1960-е гг. задача коррекции ошибок получила различные варианты решений. В 1961 г. Лес Эрнст (Les Earnest), возглавлявший исследования в этой области, решил впервые использовать алгоритм проверки орфографии со словарем из 10 000 допустимых слов. Ральф Горин — аспирант Эрнста в 1971 г. в Лаборатории искусственного интеллекта Стэнфордского университета создал первую прикладную программу проверки орфографии для английского текста — SPELL для компьютера DEC PDP-10. Алгоритм Горина выполнял поиск в списке слов правдоподобных правильных вариантов написания, отличающихся одной буквой или перестановкой соседних букв. Горин сделал программу SPELL общедоступной, как и большинство программ Стэнфордской лаборатории искусственного интеллекта, и вскоре она распространилась по всему миру. До широкого распространения персональных компьютеров оставалось ждать еще около 10 лет.

В ходе развития текстовых редакторов к середине 1980-х гг. в популярные текстовые редакторы WordStar и WordPerfect были добавлены средства проверки орфографии, с использованием разработок сторонних компаний. С английского языка поддержка была распространена на многие европейские и даже азиатские языки. Такие компании как WordPerfect стремились локализовать свое программное обеспечение для как можно большего числа национальных рынков. Несмотря на то, что затраты на разработку для ряда языков небольших рынков были убыточны.

Сегодня, в начале эпохи алгоритмов машинного обучения, или, как принято говорить в широких кругах — «искусственного интеллекта», можно наблюдать значительное усиление влияния информационных технологий на творчество человека. От рутинных задач человек в эргодических системах все более и более переходит к разрешению слабоформализованных проблем, требующих не только нестандартных подходов, гибкости ума, но и автоматизации на основе нейросетевых алгоритмов и систем (Нечаев, Евсеева, 2019).

В 2023–2024 гг. начали получать широкое распространение центральные процессоры, имеющие в своем составе специальные блоки-ускорители векторных вычислений, увеличивающие скорость работы нейросетевых алгоритмов на порядки.

Как было упомянуто, одним из прогрессивных нововведений в текстовых редакторах несколько десятилетий назад стали средства коррекции ошибок — проверки орфографии. Одним из недостатков классических алгоритмов коррекции ошибок является работа без учета контекста документа. Таким образом словам с ошибкой может быть сопоставлена некорректная замена, а некоторые из них и вовсе не будут обнаружены. Без понимания контекста невозможно, например, понять какое из слов написано с ошибкой: «сон» или «слон». Даже словосочетания «здоровый сон» и «здоровый слон» выглядят абсолютно правильными без учета темы повествования. А вот с учетом контекста оба слова могут оказаться написанными с ошибкой. Например, «В Африке мне повстречался здоровый сон» или «Здоровый слон является залогом вашего долголетия». Распространенные алгоритмы коррекции ошибок не в состоянии обнаружить ошибки такого рода, но применяя анализ, основанный на использованный обученных миллионами документов нейросетей, мы можем обнаружить логическое несоответствие слова контексту.

Есть основания полагать, что в ближайшем будущем, благодаря использованию появляющегося сейчас поколения процессоров с нейромодулями, увеличивающих скорость работы нейросетей, коррекция ошибок станет значительно более точной и чувствительной. Иными словами, новые прогрессивные алгоритмы появятся в распоряжении рядовых пользователей и, благодаря новым процессорам, смогут выполняться локально — на персональном компьютере. Примером тому может служить методика, основанная на формировании модифицированного множества ортогональных сигналов, математическими моделями которых является множество кусочно-постоянных ортогональных функций Радемахера и Уолша (Макаров, Нечаев, 2014) для обработки сложного составного многоуровневого суммарного сигнала, форма которого отображает состояние параллельного интерфейса вычислительного комплекса и его обработку псевдокорреляционными устройствами приемника.

С недавних пор существует и активно развивается множество онлайн-сервисов, использующих нейросетевые алгоритмы, которые не только позволяют корректировать грамматические ошибки и синтаксис, заменяющих редактора-человека. Работа редактора в издательстве прежде всего состоит в анализе текста на предмет его соответствия требованиям издания. Также редактор исправляет стилистические изъяны, помогает автору скорректировать стилистику текста, предлагая альтернативные варианты написания. Таким образом, алгоритмы автоматического редактора гораздо сложнее алгоритмов коррекции ошибок. Подобные сервисы могут быть полезны авторам, не имеющим возможности пользоваться услугами профессионалов, а также — при составлении текстов на иностранном (чужом) языке. Если сегодня эти службы работают в онлайн-режиме, обрабатывая текст на стороннем сервере, то в обозримом будущем можно ожидать появления таких развитых средств и на персональных компьютерах — в составе программ для редактирования текстов.

Также в будущем ожидаются изменения в верстке текста. Современные текстовые редакторы, в отличие от программ профессиональной верстки, не позволяют технически сформировать массив текста с учетом всех тонкостей общепринятых норм дизайна. Например, в русской типографике длинное тире, не стоящее в начале абзаца и не обозначающее диапазон значений, отбивается пробелами. В текстовых редакторах для этого используются обычные межсловные пробелы, в то время как профессиональный верстальщик, основываясь на традициях, формировавшихся веками, для отбивки длинного тире будет использовать узкие пробелы шириной в 2 пункта (для кегля в 10 пунктов). Существуют и другие примеры несовершенства текстовых редакторов в сравнении с программами профессиональной верстки.

По мнению авторов, уровень современных информационных технологий позволяет автоматизировать верстку в текстовых редакторах, приблизив ее к профессиональному уровню, тем самым значительно упростив создание малотиражных печатных изданий и прочих публикаций, не попадающих по разным причинам под заботливое око опытного верстальщика. Современные нейросетевые алгоритмы способны распознать тип и значение различных текстовых блоков, опираясь на их содержание, и как следствие — предложить правильный для каждого случая вариант оформления.

В 2022 г. человечество вступило в эру генеративного искусственного интеллекта, — это разновидность искусственного интеллекта, использующая генеративные модели для создания различных форм данных, например, текста, изображений, видео.

Генеративные модели обычно формируют выходные данные, соответствующие запросу пользователя. Генеративные системы искусственного интеллекта обучаются на больших массивах информации и, опираясь на полученный опыт, синтезируют новые данные (Сохина, Немченко, 2022).

Первой системой такого рода стал ChatGPT. Это выпущенный в 2022 г. чатбот, основанный на генеративном искусственном интеллекте (ИИ), разработанный комманией OpenAI. Он основан на большой языковой модели (LLM) GPT архитектуры 3.5, которая позже была обновлена до версии GPT-4. ChatGPT может формулировать ответы и суждения, подобные человеческим, а также дает возможность пользователям управлять беседой с целью получения желаемых результатов в необходимой форме. ChatGPT породил всплеск массового интереса к генеративному ИИ и как следствие — масштабным инвестициям в развитие этой отрасли.

Богатые возможности и неплохие результаты работы вызвали в обществе закономерную обеспокоенность тем, что ChatGPT и другие системы такого рода могут уже в обозримом будущем снизить спрос на человеческий интеллект и даже синтезировать дезинформацию.

К январю 2023 г. ChatGPT стал самым интенсивно развивающимся потребительским программным приложением в истории, получив всего за 2 месяца

более 100 млн. пользователей. Оценочная стоимость компании OpenAI составила к этому моменту 86 млрд. долларов. Успех ChatGPT стимулировал выпуск ряда подобных систем другими производителями, например, Gemini, Claude, Llama и др. Компания Microsoft в тесном сотрудничестве с OpenAI выпустила систему Copilot, основанную на GPT-4. В июне 2024 г. в рамках сотрудничества Apple Inc и OpenAI ChatGPT дополнил функциональность Apple Intelligence в операционных системах Apple. По статистике на июль 2024 г. сайт ChatGPT входит в число 20 самых посещаемых в мире.

Существую определённые особенности генеративного ИИ. ChatGPT уже использовался для создания вводных разделов и аннотаций к научным статьям. В нескольких статьях ChatGPT указан в качестве соавтора.

И все же реакция издательств научных журналов на ChatGPT различна. Например, Nature и JAMA Network, требуют, чтобы авторы раскрывали информацию об использовании инструментов для генерации текста и запрещают указывать в качестве соавтора языковые модели подобные ChatGPT. Также издательство Science полностью запретило использование текста, созданного при помощи генеративного ИИ, во всех своих журналах.

Известно, что ChatGPT зачастую дает правдоподобные, но неправильные или бессмысленные ответы. Эта особенность называется «галлюцинацией» и свойственна многим большим языковым моделям.

В 2023 г. испанский химик Рафаэль Луке опубликовал ряд научных работ, написанных силами ChatGPT, как позже признался ученый. В этих статьях встречаются необычные фразы свойственные большим языковым моделям. По мнению многих ученых время для использования ChatGPT в академической среде для написания рецензий и учебных пособий еще не наступило по причине спорадического проявления в их творчестве «галлюцинаций», сильно искажающих смысл повествования. Например, в рецензиях, созданных ChatGPT могут попадаться упоминания о несуществующих исследованиях, «придуманных» нейросетью.

Несмотря на очевидные проблемы современных систем генеративного ИИ, следует признать, что за пару лет с начала своего широкого распространения они значительно прогрессировали и продолжают развиваться. Надо полагать, что в обозримом будущем создание текста во многих областях человеческой деятельности будет успешно и безошибочно решаться силами генеративного ИИ.

Хорошо это или плохо — вопрос во многом философский. Даже среди специалистов отрасли нет единого мнения. В марте 2024 г. Илон Маск и Стив Возняк опубликовали открытое письмо с призывом остановить обучение больших нейросетей: «Мы призываем все лаборатории, исследующие искусственный интеллект, немедленно прекратить хотя бы на шесть месяцев тренировки систем мощнее GPT-4. Эта пауза должна быть публичной и действительной. Если же подобная приостановка не может быть сделана быстро, правительства

государств должны вмешаться и ввести мораторий» (Маск и Возняк…, Электр. ресурс). Документ подписало более тысячи специалистов отрасли и руководителей профильных компаний. По мнению подписантов, обучение больших систем ИИ необходимо приостановить до появления общих правил безопасности, а сама пауза должна находиться под контролем группы независимых экспертов.

Как бы то ни было, следует признать, что технологии работы с текстовой информацией преодолели путь от простых механических станочных оттисков до частичного вытеснения человека-автора из творческого процесса. Но, несмотря на то, что в этом контексте высокая динамика развития искусственного интеллекта вызывает тревогу, нам придется жить в этом новом мире, учиться использовать новые возможности преимущественно во благо человечеству. Подобно тому, как нож чаще используют не для победы в споре, а для приготовления еды.

## *СПИСОК ЛИТЕРАТУРЫ*

Костина А. В. (2022) Массовое сознание: от индустриального — к информационному обществу // Ученый совет. № 12. — С. 763–769.

Макаров В. Ф., Нечаев Д. Ю. (2014) Устранение избыточности в системах ортогонального кодирования // Безопасность информационных технологий. Т. 21, № 2. — С. 54–59.

Маск и Возняк попросили приостановить разработку систем ИИ, превосходящих по мощности GPT-4 (2024) //Интерфакс [Электронный ресурс] URL: https://www.interfax.ru/world/893528 (дата обращения: 27.10.2024).

Нечаев Д. Ю., Евсеева А. Ю. (2019) Интеллектуальная виртуальная образовательная среда как средство достижения образовательных целей в компетентностной парадигме // Высшее образование для XXI века: роль гуманитарного образования в контексте технологических и социокультурных изменений: XV Международная научная конференция. Доклады и материалы. В 2-х частях, Москва, 14–16 ноября 2019 года / Под общей редакцией И. М. Ильинского. Том Часть 2. — М: МосГУ, С. 490–495.

Основы информационных технологий (2010): Учебное пособие / Киреева Г. И., Курушин В. Д., Мосягин А. Б., Нечаев Д. Ю., Чекмарев Ю. В., М.: ДМК Пресс, 273 с.

Сохина С. А., Немченко С. А. (2022) Машинное обучение. Методы машинного обучения // Современная наука в условиях модернизационных процессов: проблемы, реалии, перспективы: Сборник научных статей по материалам V Международной научно-практической конференции, Уфа, 30 апреля 2022 года. — Уфа: Вестник науки, С. 165–168.

Шомракова И. А., Баренбаум И. Е. (2008) Всеобщая история книги, СПб.: Профессия — СПбГУКИ, 392 с.: ил.

***Гаврилова Ольга Викторовна,*** доцент кафедры прикладной информатики и статистики Московского гуманитарного университета. Адрес: 111395, Россия, г. Москва, ул. Юности, д. 5. Тел.: +7 (915) 322 98 98, Эл. адрес: gavrilova.ov@gmail.com

***Gavrilova Olga Viktorovna,*** Associate Professor of the Department of Applied Informatics and Statistics, Moscow University for the Humanities, Address: 111395, Russia, Moscow, Yunosti str., 5. Tel.: +7 (915) 322 98 98, Email: gavrilova.ov@gmail.com

***Зубковский Павел Алексеевич,*** старший аналитик IPSOS. Тел.: +7 903 006 11 33. Эл. адрес: pavel@zubkovsky.ru

***Zubkovsky Pavel Alekseevich,*** Senior Analyst at IPSOS, Tel.: +7 903 006 11 33, Email: pavel@zubkovsky.ru

***Для цитирования:***

Гаврилова О.В., Зубковский П.А. Работа с текстом: прошлое, настоящее, будущее. № 5. С. 14–22. DOI: https://www.doi.org/10.17805/trudy.2024.5.3