УДК 004.9, 81’32
DOI 10.17726/phillT.2024.2.4



# Эволюция методов обработки естественного языка

***Беседина Анастасия Юрьевна,***
*студентка 1-го курса магистратуры*
*кафедры теории и практики перевода*
*Института переводоведения, русистики и многоязычия,*
*Пятигорский государственный университет*
*Пятигорск, Россия*

*nastya.besedina.02@mail.ru*

***Аннотация.*** Обработка естественного языка (Natural language processing, NLP) претерпела значительные изменения в своей методике, что отражает прогресс в области вычислительных технологий и когнитивных исследований. В данном обзоре мы осветим ключевые моменты эволюции методов анализа естественного языка. В статье затрагивается тема первых разработанных систем NLP, приведены обоснования причин сложности некоторых обрабатываемых текстов и возможной глубины анализа. Кроме того, здесь описаны не только методы NLP до и после GPT-революции, но и современные тенденции и перспективы в сфере обработки естественного языка. Статья позволяет проследить, как менялось представление о тексте на естественном языке в ходе развития компьютерных методов анализа, а также разобраться, чем является текст в зеркале обработки естественного языка, что в действительности является предметом исследования обработки естественного языка и чего нельзя увидеть глазами простого исследователя, не использующего методы NLP.

**Ключевые слова:** обработка естественного языка; анализ текста; виртуальные ассистенты; трансформеры; искусственный интеллект; нейронные сети.

# Evolution of natural language processing methods

***Besedina Anastasia Yuryevna,***
*first-year graduate student,*
*Department of Theory and Practice of Translation*
*Institute of Translation Studies, Russian Studies and Multilingualism*
*Pyatigorsk State University*
*Pyatigorsk, Russia*

*nastya.besedina.02@mail.ru*

**Abstract.** Natural language processing (NLP) has undergone significant changes in its methods, reflecting advances in computing technology and cognitive research. This article reviews the key stages of the evolution of natural language processing methods. The article touches on the topic of the first NLP systems developed, provides justification for the reasons for the complexity of some processed texts and the possible depth of analysis. In addition, it describes not only NLP methods before and after the GPT revolution, but also current trends and prospects in the field of natural language processing. The article allows us to trace how the idea of natural language text has changed during the development of computer analysis methods, as well as to understand what text is in the mirror of natural language processing, what is really the subject of natural language processing research and what cannot be seen through the eyes of a simple researcher who does not use NLP methods.

**Keywords:** natural language processing; text analysis; virtual assistants; transformers; artificial intelligence; neural networks.

## Введение

Исследование и обработка естественного языка (Natural language processing, NLP) представляет собой многоаспектную научную дисциплину, связывающую лингвистику, информатику, когнитивные науки и области искусственного интеллекта. Специалисты этой сферы занимаются созданием вычислительных систем, которые могут осмысливать, анализировать и производить тексты на человеческом языке. Эволюция методов NLP связана с развитием вычислительных мощностей, сменой парадигм и появлением новых теоретических подходов в области когнитивных наук и анализа данных. Данная статья представляет исторический обзор эволюции методов NLP, выделяя ключевые этапы и изменения восприятия явления «текст».

Можно сказать, что обработка естественного языка — это изучение вычислительных методов для автоматического анализа текста и извлечения значимой информации для последующего анализа. Сложность анализа текста «заключается в том, что текст эллиптичен, неполон и насквозь пронизан умолчаниями» [1, с. 106]. **Актуальность** темы статьи обусловлена высоким ростом значимости обработки естественного языка в различных областях деятельности человека. Изучение эволюции методов NLP способно помочь разобраться в работе этой системы, начиная с ее

истоков, и дать достаточно полное представление о развитии методов в историческом срезе.

**Цель** данной статьи состоит в рассмотрении эволюции методов NLP с исторической точки зрения. **Задача** — проследить, как менялось представление о тексте на естественном языке в ходе развития компьютерных методов анализа.

«Обработка естественного языка — это наука о проектировании методов и алгоритмов, которые принимают или порождают неструктурированные данные естественного языка» [2, с. 19]. Эта сфера знаний занимается взаимодействием между компьютерами и человеческим языком, где главная задача заключается в обеспечении компьютерам способности понимать, анализировать и воспроизводить человеческий язык так же, как это делает человек. В связи с этим, сюда входят такие задачи, как понимание значения (способность извлекать значение из текста, речи или других форм человеческого языка), анализ структуры (распознавание грамматической структуры и синтаксиса языка, включая части речи и построение предложений), создание языка, подобного человеческому (создание естественного, связного и грамматически правильного текста или речи).

Н. Н. Леонтьева описывает различные виды структур, необходимых для понимания текста: лингвистические структуры в предложениях текста, что можно назвать локальным пониманием; семантические сети в целом тексте, что представляет глобальное размытое понимание; информационные структуры в тексте в целом, что дает глобальное обобщенное понимание; и структуры баз данных и знаний, что обеспечивает выборочное специальное понимание [3]. NLP нацелено на преодоление разрыва между человеческим общением и машинным пониманием, способствуя беспрепятственному взаимодействию между людьми и технологиями. При этом, «помимо проблем, связанных с обработкой неоднозначных и вариативных входных данных в системе с плохо определенными и отсутствующими наборами правил», у естественного языка «есть и дополнительные свойства, которые еще больше затрудняют разработку вычислительных подходов на основе машинного обучения: дискретность, композиционность и разреженность» [2, с. 19].

Эволюция методов обработки естественного языка в историческом контексте тесно связана с этапами зарождения и прогресса

машинного перевода, распознавания речи и искусственного интеллекта.

В XVII веке, который ознаменован вкладом ведущих философов, зародился путь развития машинного перевода. Рене Декарт предложил в 1629 году универсальный язык, который содержал эквивалентные идеи на разных языках, однако использующих один символ. В 30-е годы прошлого века были выданы первые патенты на устройства для перевода текстов. Стоит отметить автоматический двуязычный словарь Жоржа Арцруни, который функционировал с помощью бумажной ленты. Питер Троянский предложил концепцию, включающую в себя двуязычный словарь и метод определения грамматических функций между языками. Он основывался на системе языка эсперанто. В 1950 году Алан Тьюринг в своей знаменитой работе «Вычислительная техника и интеллект» представил «тест Тьюринга» как мерило разумности. Этот критерий оценивает способность компьютерной программы на достаточно высоком уровне имитировать человека в процессе текстового разговора с человеком-судьей, настолько, чтобы последний не мог однозначно определить, с кем он общается — с программой или с человеком. Семь лет спустя труд Ноама Хомского «Синтаксические структуры» потряс лингвистику, введя концепцию «универсальной грамматики», представляющую собой систему синтаксических правил. Все вышеперечисленное стало предпосылкой к дальнейшему возникновению новых и изменению существующих на тот момент методов обработки естественного языка.

### Первый этап развития методов обработки естественного языка (с 1950-х гг. по 1970-е гг.)

Самые первые разработки систем обработки естественного языка основывались на символьном подходе. Важно понимать, что в то время текст рассматривался скорее как последовательность символов, которую можно анализировать посредством определенных правил и формальных грамматик. Именно тогда стали разрабатываться такие формальные языки или системы описания синтаксиса, как, например, Форма Бэкуса — Наура (БНФ) (Backus-Naur Form — BNF), которые описывали синтаксис языка и анализировали грамматическую структуру предложений, в процессе чего одни синтаксические категории последовательно определялись через другие. Сложность обрабатываемых текстов и глубину ана-

лиза, прежде всего, ограничивало то, что вычислительные мощности были низкими. Чтобы иметь полное представление о развитии обработки естественного языка, необходимо принять во внимание систему ELIZA, которая стала важнейшим достижением того периода, однако, в действительности, демонстрировала ограниченное понимание языка и опиралась на ключевые слова, заготовки, шаблоны, а не на истинное понимание семантики. Ограниченность вычислительных мощностей не давала возможность обрабатывать большие объемы текста и учитывать сложные лингвистические явления, поскольку внимание в большей степени уделялось синтаксическому анализу, а семантический анализ был очень простым. Кроме того, к подходам, основанным на правилах и символьной обработке информации, относятся системы на основе древовидных структур, регулярных выражений и т.д. Специалисты отмечают: «Примерно с 1960 по 1985 год в большинстве областей лингвистики, психологии, искусственного интеллекта и обработки естественного языка полностью доминировал рационалистический подход. Рационалистический подход характеризуется верой в то, что значительная часть знаний, хранящихся в человеческом сознании, не извлекается органами чувств, а фиксируется заранее, предположительно путем генетического наследования» [4, с. 4]. Можно сделать вывод, что сам текст в этой парадигме — последовательность символов, поддающаяся формализации и разбору по заранее заданным правилам. В таком случае «невидимым» оставалось тонкое понимание контекста, многозначности слов и нюансов человеческой речи.

### Второй этап развития методов обработки естественного языка (с 1980-х гг. по 1990-е гг.)

По мере роста вычислительных мощностей и накопления больших корпусов текстовых данных стало возможным использование статистических методов. Вместо строго определенных правил, статистические модели теперь обучались на корпусах текстов, выявляя вероятностные связи между словами и фразами. В мире анализа последовательностей возникли новые технологии — скрытые марковские модели (Hidden Markov Model — HMM) и методы n-грамм. Они способны предсказывать следующее слово в предложении, опираясь на предыдущие слова. N-грамма представляет собой последовательность из n элементов, которая может быть

звуковой, слоговой, словесной или буквенной. Такие методы, как n-граммы, скрытые марковские модели и статистический машинный перевод, открыли новые возможности для анализа текстовых данных и помогли усовершенствовать точность анализа текста и решение более сложных задач, как, например, разбор предложений и частотный анализ. Однако статистические модели зачастую не могли в достаточной мере обработать контекст и семантические отношения между словами, что приводило к ошибкам в анализе многозначных слов и сложных предложений. Закономерно, что текст стал рассматриваться как вероятностное распределение слов и фраз, где важность отдельных элементов определяется их частотой и соседством. Здесь соответственно «невидимым» оставались более глубокие семантические отношения, неявные связи между словами и сложные лингвистические явления.

**Третий этап развития методов обработки естественного языка (с 2000-х гг. по настоящее время)**

Развитие глубокого обучения стало настоящим революционным прорывом в обработке естественного языка. Основой большинства современных систем NLP являются нейросети и глубокое обучение, поскольку данные технологии дают возможность построить модель сложных зависимостей данных и извлечь скрытые закономерности. Благодаря нейронным сетям, особенно рекуррентным нейронным сетям (Recurrent Neural Network — RNN), механизмам долгосрочной краткосрочной памяти (Long Short Term Memory — LSTM), а также трансформерам, стало возможным осуществлять более глубокий анализ и изменение текста. Обратимся к мнению современных исследователей: «Рекуррентные сети являются весьма выразительными моделями для последовательностей и являются, пожалуй, самым полезным, что могут предложить нейронные сети обработке языков. Они позволяют отказаться от марковского предположения, преобладавшего в NLP в течение нескольких десятилетий, и проектировать модели, в которых условиями могут быть целые предложения. При этом они могут при необходимости учитывать порядок слов и не подвержены проблемам статистического оценивания, проистекающим из разреженности данных. Эта возможность дает заметный выигрыш в языковом моделировании — задаче о предсказании вероятности следующего слова в последовательности (или, что то же самое, вероятности

предложения), — которое является краеугольным камнем многих приложений NLP» [2, с. 22]. Векторные представления слов (word embeddings) позволили кодировать семантическую информацию в многомерном пространстве, где схожие по значению слова располагаются близко друг к другу. Механизм внимания в трансформерах позволил моделям фокусироваться на важных частях текста, учитывая контекст и взаимосвязи между различными его частями. Предварительно обученные модели, такие как BERT (Bidirectional encoder representations from transformers), XLNet (авторегрессионная языковая модель, которая выдает на выходе вероятность совместной встречаемости последовательности токенов) и GPT (Generative Pre-trained Transformer), достигли невероятных успехов в различных задачах обработки естественного языка от машинного перевода до генерации текста. Несмотря на то, что нейросети используют большие вычислительные ресурсы и огромные объемы тренировочных данных, проблема смещения и субъективности в обучающих данных также остается актуальной. На текущем этапе развития методов обработки естественного языка текст рассматривается как многомерное семантическое пространство, где связи между его элементами закодированы в сложной структуре нейронной сети. В тексте «невидимым» до сих пор еще остается полное понимание человеческого языка с его нюансами, амбивалентностью и неявным значением.

Рассматривая современные методы обработки естественного языка, важно иметь представление об основных направлениях применения искусственного интеллекта (ИИ) в NLP как важнейшего результата GPT-революции. Одним из ключевых направлений использования искусственного интеллекта в сфере обработки текстовых данных (NLP) является автоматический перевод текстов. В наше время активно применяются переводческие системы на базе искусственного интеллекта, такие как Google Translate и DeepL, в которых для осуществления перевода используются нейросетевые технологии. Эти системы обладают способностью учитывать контекст и семантику слов в тексте, что значительно повышает качество перевода по сравнению с более примитивными статистическими методами. Более того, вспомогательные программы Google Assistant, Siri и Alexa также применяют искусственный интеллект для того, чтобы корректно распознавать и обрабатывать команды пользователей, выраженные на естественном

языке. Эти системы способны отвечать на вопросы пользователей, помогать в организации встреч и заказов, а также поддерживать диалог. Другой важной особенностью является то, что при анализе текстов в различных приложениях стали активно использоваться нейросети. В качестве примера можно привести систему, способную самостоятельно распределять тексты по положительной или отрицательной категории мнений, тональности на основе анализа используемых слов и фраз. Эта технология находит применение, в том числе, в исследованиях рынка и для осуществления оперативного контроля обратной связи от клиентов и динамики активности в социальных сетях. Искусственный интеллект также активно применяется для разделения текстов на группы, выделения ключевых слов и понятий, а также для семантического анализа. Важным направлением использования ИИ в области обработки естественного языка является распознавание речи. Сервисы Google и Speech-to-Text используют нейронные сети для преобразования звука и аудиофайлов в текст. Кроме того, существуют алгоритмы, способные автоматически создавать краткие выжимки даже для очень длинных текстов, выделяя ключевую информацию и сохраняя логическую связь и смысл исходного содержимого. «Нейронные сети дают эффективный механизм обучения, чрезвычайно привлекательный для использования в задачах обработки естественного языка, — утверждают специалисты. — Главный компонент языковой нейронной сети — слой погружения, т.е. отображение дискретных символов на непрерывные векторы в пространстве сравнительно небольшой размерности. В результате погружения слова преобразуются из изолированных дискретных символов в математические объекты, над которыми можно производить различные действия» [2, с. 21].

При обработке текста на естественном языке исследователь обладает всеми необходимыми инструментами и алгоритмами для классификации текста, извлечения информации, проведения тематического моделирования и даже анализа настроений.

Обратимся к методам обработки естественного языка для более детального рассмотрения. Когда возникает необходимость провести классификацию текста, он распределяется по заранее определенным категориям в зависимости от их тематики и наполнения. Следовательно, модели классификации на базе методов обработки естественного языка имеют возможность автоматически распреде-

лять и группировать документы относительно их тематики и тона. С помощью извлечения информации исследователь, как правило, приобретает информацию, обладающую четкой структурой, пусть до этого она и являла собой беспорядочные данные. Распознавание и извлечение определенных событий или отношений способствуют эффективному использованию поиска информации и построению граф-схем знаний. Автоматическое определение тем на основе текстовых данных и распределение документов по темам позволяют получать сгруппированные данные и анализ тенденций. Согласимся с мнением, что «количество текстов на естественном языке, доступных в электронном виде, поистине ошеломляет и увеличивается с каждым днем. Однако сложность естественного языка может сильно затруднить доступ к информации, содержащейся в этом тексте. Современный уровень развития НЛП все еще далек от того, чтобы создавать универсальные представления смысла из неограниченного текста» [5, с. 261].

Рассматривая современные тенденции в области обработки естественного языка и ее перспективы, следует отметить, что сейчас большинство исследований в данной сфере нацелено на многоязычное NLP. К нему относится разработка моделей, способных работать с разными языками. Также существует необходимость в разработке подходов к обработке информации на языках, для которых имеется недостаточно данных. Это включает в себя создание алгоритмов для анализа процессов принятия решений в сложных нейронных сетях. Важны также интеграция NLP с различными направлениями искусственного интеллекта, объединение NLP с компьютерным зрением и робототехникой для формирования более совершенных и умных систем.

Отвечая на вопрос, чего нельзя увидеть глазами простого исследователя, становится понятно, что простой анализ текста, не использующий методы обработки естественного языка, не способен выявить множество взаимосвязей, скрытых от человека при поверхностном анализе, среди которых тонкие смысловые связи между словами, не явные из синтаксиса, зависимость значения слов от контекста и окружающих слов, неявный смысл, передаваемый не только словами, но и интонацией, жестами, и различные значения одного и того же слова в разных контекстах и эмоциональная окраска. Для полноценной работы система анализа текста «должна иметь возможность проанализировать тест, поданный пользовате-

лем 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 просинтезировать свой отклик на естественном языке» [1, с. 106].

Таким образом, можно сделать вывод, что в контексте развития технологий представления о тексте постоянно менялись. С каждым новым этапом развития методов обработки естественного языка люди все больше начинали интересоваться своим языком и, как следствие, текстом, его беспрерывными изменениями и пытались проводить сравнения в историческом срезе. «Естественный язык находится в постоянном движении, адаптируясь к меняющемуся миру, создавая названия и слова для обозначения новых вещей, новых людей и новых концепций» [4, с. 309], следовательно, текст в зеркале обработки естественного языка предстает не просто набором слов, а сложной структурой, отражающей семантику, синтаксис, прагматику и контекст. Именно скрытые взаимосвязи между этими элементами, недоступные простому человеческому наблюдению, являются предметом исследования NLP.

**Заключение**

1. При первых попытках разработки систем NLP, с 1950-х по 1970-е гг., методы основывались преимущественно на символьном подходе, а текст представлялся последовательностью символов, подлежащей анализу с помощью формальных грамматик и правил. Однако все еще предстояло искать пути к тонкому пониманию контекста, многозначности слов и нюансов человеческой речи.

2. На втором этапе развития методов обработки естественного языка, с 1980-х по 1990-е гг., текст стал рассматриваться как вероятностное распределение слов и фраз, где важность отдельных элементов определялась их частотой и соседством. Глубокие семантические отношения, неявные связи между словами и сложные лингвистические явления сохраняли свою научную непостижимость.

3. Третий этап развития методов обработки естественного языка, с 2000-х гг. по настоящее время, отличается тем, что текст трактуется как многомерное семантическое пространство, в котором связи между элементами закодированы в сложной структуре

нейронной сети. Однако результат оставляет задачу полного понимания человеческого языка с его нюансами, амбивалентностью и неявным значением для будущих исследований.

4. Главной движущей силой существующих на данный момент систем обработки естественного языка можно назвать нейросети, трансформеры и глубокое обучение. К преимуществам таких систем относятся улучшение качества и точности обработки и анализа, автоматизация рутины, персонализация, к вызовам — проблемы с интерпретацией и объяснимостью, требования к данным и ресурсам, этические и социальные аспекты.

5. Анализ текста без использования методов NLP не способен выявить большинство скрытых взаимосвязей, среди которых семантические отношения, контекстуальное значение, прагматика и многозначность.

6. В зеркале обработки естественного языка на данный момент текст является сложной структурой, отражающей семантику, синтаксис, прагматику и контекст.

Эволюция NLP показывает, как постепенный рост вычислительных мощностей и развитие новых методов позволяют все глубже проникать в сложную структуру человеческого языка. Будущие прорывы в NLP будут определяться развитием новых алгоритмов и увеличением доступности вычислительных ресурсов и больших языковых моделей. Данная статья дает полноценную картину этапов эволюции обработки естественного языка в своих методах и позволяет проследить взаимосвязь развития методов анализа и представления о тексте на естественном языке.

***Литература***

1. Большакова Е. И., Клышинский Э. С., Ландэ Д. В., Носков А. А., Пескова О. В., Ягунова Е. В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с. (Bolshakova E. I., Klyshinsky E. S., Lande D. V., Noskov A. A., Peskova O. V., Yagunova E. V. Automatic text processing in natural language and computational linguistics: textbook. the manual. M.: MIEM, 2011. 272 p.)
2. Гольдберг Й. Нейросетевые методы в обработке естественного языка. Пер. с англ. А. А. Слинкина. М.: ДМК Пресс, 2019. 282 с. (Goldberg J. Neural network methods in natural language processing. Translated from English by A. A. Slinkin. M.: DMK Press, 2019. 282 p.)
3. Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие. М.: ИЦ «Академия», 2006. 304 с. (Leontieva N. N. Automatic understanding of texts: systems, models,

resources: a textbook. M.: IC “Academia”, 2006. 304 p.)
4. Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing, 8, [print.]-e изд. Cambridge, Massachusetts: The MIT Press, 2005. 720 p.
5. Bird S., Klein E., Loper E. Natural language processing with Python. 1st ed-e изд. Beijing; Cambridge [Mass.]: O’Reilly, 2009. 479 c.
6. Тюрина Д. А., Пальмов С. В. Применение нейронных сетей в обработке естественного языка // Журнал прикладных исследований. 2023. № 7. С. 158-162. (Tyurina D. A., Palmov S. V. Application of neural networks in natural language processing // Journal of Applied Research. 2023. № 7. P. 158-162.).
7. Jackson P., Moulinier I. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization // Computational Linguistics. 2003. № 29. P. 510-511.
8. Jelinek F. Statistical methods for speech recognition. Cambridge (GB): MIT press, 1998. 305 p.
9. Jurafsky D., Martin J. H. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2023. 636 p.
10. Kumar L. A., Renukay D. K. Deep learning approach for natural language processing, speech, and computer vision: techniques and use cases. First edition-e изд., Boca Raton, FL: CRC Press, 2023. 246 p.
11. Garg M., Kumar S. Natural language processing and information retrieval: principles and applications. London, New York: CRC Press, 2024. 271 p.
12. Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. Cambridge: Cambridge university press, 2008. 528 p.
13. Everaert M. B. H., Huybregts M. A. C., Chomsky N., Berwick R. C., Bolhuis J. J. Structures, not strings: Linguistics as part of the cognitive sciences // Trends in Cognitive Sciences. 2015. № 19(12). P. 729-743.
14. Mihalcea R., Radev D. Graph-based natural language processing and information retrieval. Cambridge; New York: Cambridge University Press, 2011. 192 p.
15. Rothman D. Transformers for natural language processing and computer vision: explore generative AI and large language models with Hugging Face, ChatGPT, GPT-4V, and DALL-E3. Third edition, revised publication Sept 2024-e изд., Birmingham, UK Mumbai: Packt Publishing Ltd, 2024. 693 p.