Научная статья
УДК 004.032.26.056+004.415.2:005.6
ББК 32.813+32.972.53
Ч 25
DOI: 10.53598/2410-3225-2022-4-311-81-89

Подход к решению проблемы контроля качества в сфере услуг на основе построения системы интеллектуального анализа данных

(*Рецензирована*)

Вера Аркадьевна Частикова[1], Виктория Геннадьевна Гуляй[2], Сергей Анатольевич Жерлицын[3]

[1–3] *Кубанский государственный технологический университет, Краснодар, Россия*
[1, 2] *chastikova_va@mail.ru*
[3] *kpytooooo@gmail.com*

Аннотация. *В статье рассмотрен новый подход к контролю качества в сфере услуг путем внедрения программного обеспечения на основе интеллектуальных систем, включающих, прежде всего, методы обработки и анализа естественного языка с применением регуляризации данных. Приводятся результаты проведенного сравнительного анализа нескольких, наиболее часто применяемых, нейронных сетей-трансформеров: BERT, GPT и XLNet. В ходе исследования было выявлено, что алгоритм BERT благодаря двунаправленному обучению больше подходит для обработки входной информации; модель GPT за счет авторегрессии – для генерации ответов системы; метод XLNet благодаря механизму PLM – для языкового моделирования. Было выявлено, что каждый из рассмотренных алгоритмов имеет свои преимущества и недостатки, поэтому для достижения оптимальных результатов следует использовать программные комплексы, базирующиеся на нескольких алгоритмах-трансформерах.*

Ключевые слова: *искусственный интеллект, анализ естественного языка, GPT, BERT, XLNet, сфера услуг, оценка контроля качества*

Original Research Paper

The approach to solving the problem of quality control in the service sector based on the construction of a data mining system

Vera A. Chastikova[1], Victoriya G. Gulyay[2], Sergey A. Zherlitsyn[3]

[1–3] *Kuban State University of Technology, Krasnodar, Russia*
[1, 2] *chastikova_va@mail.ru*
[3] *kpytooooo@gmail.com*

Abstract. *This article discusses a new approach to quality control in the service sector by implementing software based on intelligent systems, including, first of all, methods of processing and analyzing natural language using data regularization. The article also presents the results of a comparative analysis of several, the most commonly used, neural networks-transformers: BERT, GPT and XLNet. In the course of the study, we have revealed that the BERT algorithm, thanks to bidirectional learning, is more suitable for processing input information; the GPT model, due to autoregression, for generating system responses; the XLNet method, thanks to the PLM mechanism, for language modeling. The research shows that each of the considered algorithms has its advantages and disadvantages, therefore, in order to achieve optimal results, we should use software complexes based on several transformer algorithms.*

Keywords*: artificial intelligence, natural language analysis, GPT, BERT, XLNet, service sector, quality control assessment*

Введение

В настоящее время сфера оказания услуг является одной из самых быстро развивающихся отраслей экономики как в нашей стране, так и за рубежом. Область предоставления услуг непосредственно влияет на уровень социального благополучия населения, так как затрагивает все ключевые сферы жизни людей.

Для поддержания данной области на должном уровне необходимо регулярно проводить оценочные мероприятия с целью выявления возможных недостатков и их своевременного устранения. Также оценочный анализ качества осуществляемой работы представляет коммерческий интерес для владельцев и учредителей компаний, так как уровень оказываемых услуг напрямую связан с количеством привлекаемых клиентов, а, соответственно, и с доходом компании.

1. Необходимость внедрения системы оценивания

Сфера услуг – область экономики, включающая в себя все виды коммерческих и некоммерческих услуг, в том числе отрасли здравоохранения, образования, торговли и т.д. [1]. Таким образом, каждый человек практически ежедневно сталкивается с потребностью в предоставлении ему тех или иных услуг, становясь при этом клиентом различных организаций. Однако рынок услуг включает в себе и такие отрасли, которые требуют наличия специальных знаний и навыков. Компании, стремясь повысить сервисное обслуживание, предусматривают консультации клиентов в специфических вопросах за счет пополнения штата работников менеджерами и консультантами, а также за счет внедрения колл-центров и служб поддержки для связи с клиентами с помощью различных средств коммуникаций (телефонные звонки, онлайн-чаты и т.д.).

С целью повышения качества обслуживания необходимо контролировать работу менеджеров, консультантов и служб поддержки. В частности, нужно отслеживать профессиональную подготовку работников, корректность их ответов, способность решить ту или иную проблему, помочь клиенту с выбором, а также тактичность и доброжелательность в общении. Для контроля качества можно выделить отдельного сотрудника, в обязанности которого будет входить проверка корректности работы консультирующих служб организации. Однако такой способ контроля неэффективен, так как один человек не способен обработать огромное количество звонков, поступающих в колл-цент в течение дня. Выходом из этой ситуации может стать повышение числа контролирующих органов, однако это влечет повышение затрат организации в виде заработной платы работников, обеспечения их оборудованными рабочими местами и т.д. Также важно отметить, что такая оценка будет являться субъективной, так как она будет зависеть от личностных качеств сотрудника, осуществляющего контроль.

Для получения полной и объективной оценки работы консультирующих служб организации необходимо автоматизировать процесс анализа данных и оценивания полученных результатов. Это обусловлено наличием следующих факторов, связанных с процедурой оценивания:

– для получения полного представления о качестве оказываемых услуг требуется иметь большую базу клиентских отзывов;

– для обеспечения объективности оценивания необходимо при анализе данных исключить такие человеческие факторы, как эмоциональность, симпатия/антипатия, невнимательность, отсутствие опыта в решении аналогичных задач и т.д.;

– для поддержания актуальности оценки, получаемой в ходе анализа, имеющиеся данные должны регулярно обновляться и варьироваться в зависимости от происходящих изменений в качестве предоставляемых услуг.

Перечисленные факторы указывают на необходимость создания интеллектуальной самообучающейся системы [2, 3], позволяющей получать объективную оценку ка-

чества оказываемых услуг какой-либо организацией.

2. Предлагаемый подход

Предлагаемый подход решения поставленной задачи предусматривает внедрение в организации интеллектуальной системы, базирующейся на алгоритмах анализа естественного языка с применением регуляризации данных. Обобщенная схема предлагаемого решения представлена на рисунке 1.

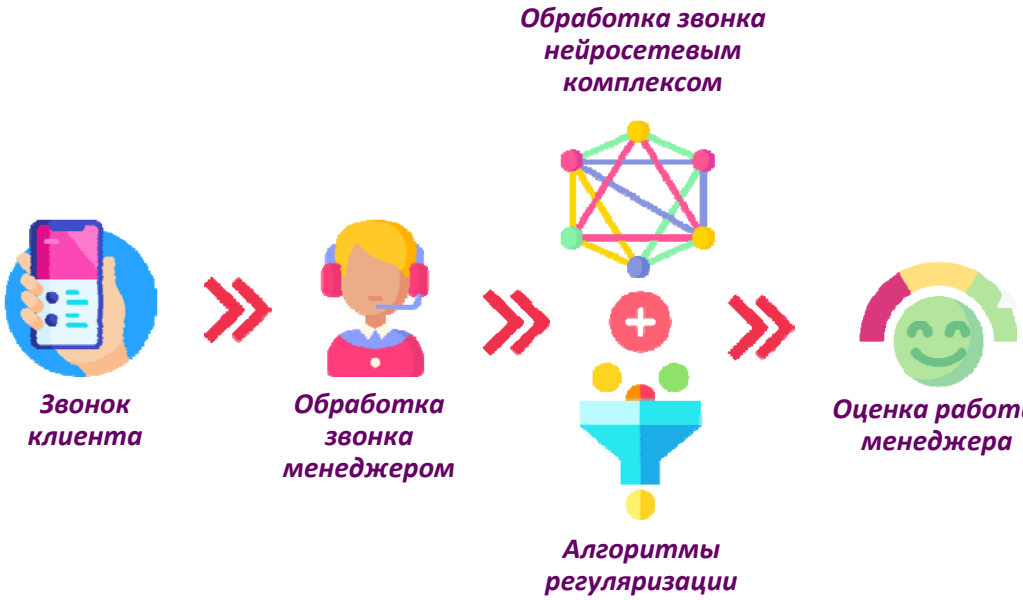

Рис. 1. Обобщенная схема предлагаемого решения

Fig. 1. Generalized scheme of the proposed solution

Система анализа и оценивания работы служб поддержки (далее – система) предполагает следующий алгоритм работы:

1. Клиент обращается в службу поддержки по интересующему его вопросу (жалобе, предложению) с помощью любых электронных средств коммуникаций;
2. Сотрудник службы поддержки обрабатывает запрос клиента;
3. Комплекс, базирующийся на алгоритмах искусственного интеллекта, контролирует работу менеджера и при необходимости выводит нужную информацию в виде подсказок;
4. По итогу проведенного анализа система выводит обобщенную оценку работы менеджера, на основе которой руководитель уже может делать выводы о компетентности сотрудника.

На основе описанного алгоритма система должна решать следующие задачи:

– распознавание устной речи из телефонных разговоров и письменной из чатов мессенджеров и писем электронной почты;

– анализ семантической нагрузки и эмоциональной окраски диалога;

– прогнозирование корректного решения возникшей проблемы.

Для решения вышеописанных задач необходимо применение наиболее современных алгоритмов обработки естественного языка, базирующихся на искусственных нейронных сетях с архитектурой Трансформер.

3. Модели Трансформеров

До недавнего времени для решения задач в сфере обработки естественного языка чаще всего применялись рекуррентные нейронные сети. Архитектура рекуррентных нейронных систем весьма эффективна в системах, нацеленных на решение задач с *по-*

следовательной обработкой данных, например, задач машинного перевода. Такие модели обрабатывают слова текста последовательно, в том порядке, в котором они появляются в контексте, по одному слову однократно, без повторов. В результате такие системы трудно распараллелить, и они плохо сохраняют контекстуальные связи между вводимыми длинными текстами. Тем не менее, с появлением более сложных систем, базирующихся на анализе естественного языка, возникла необходимость в *параллельной* обработке данных.

Так, в 2017 году была предложена концепция нейронной сети-трансформера [4], позволяющей производить параллельные вычисления при обработке данных и благодаря этому сохранять контекстные связи между словами. Это стало возможным благодаря механизму Attention [5]. Данный механизм позволяет модели одновременно обрабатывать информацию из разных подпространств, на которые предварительно разбивается текст. Кодировщик механизма Attention состоит из 6 идентичных слоев, каждый из которых имеет два подслоя. Один из слоев представляет собой нейронную сеть прямого распространения, второй – алгоритм Self-Attention [5]. Декодер механизма Attention также состоит из 6 идентичных слоев. Однако в дополнение к двум подуровням в каждом слое декодера присутствует третий подуровень с алгоритмом Multi-Head Attention, применяемым к выходным данным стека кодера [5].

На сегодняшний день существует несколько моделей с архитектурой Трансформер; наиболее распространенными являются BERT, GPT и XLNet.

4. Алгоритм GPT

GPT-модель (generative pre-training model) представляет собой языковую модель, носящую авторегрессионный характер и обладающую односторонним алгоритмом обучения (рис. 2).

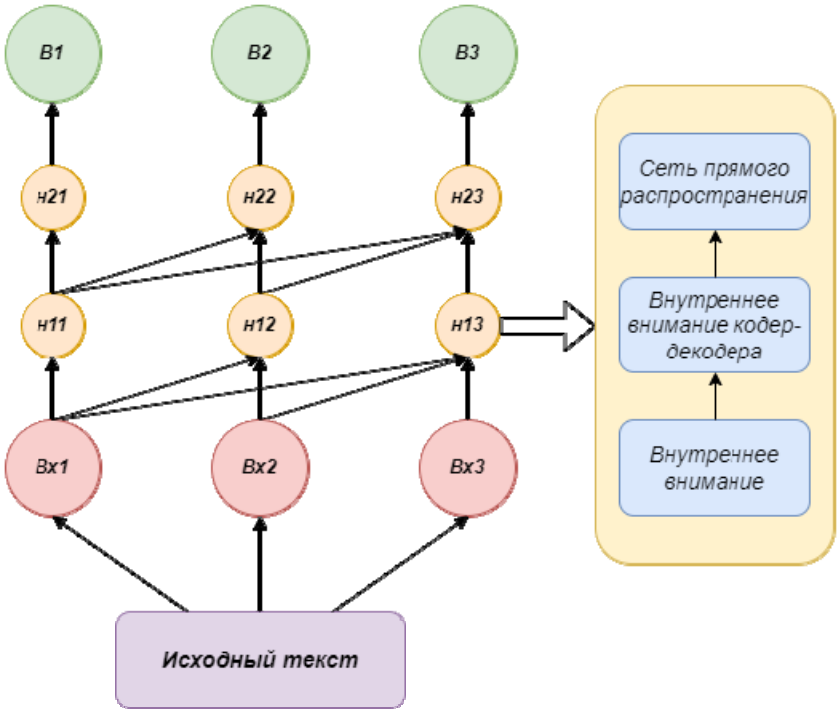

Рис. 2. Схема алгоритма GPT

Fig. 2. Scheme of the GPT algorithm

Алгоритм GPT первого поколения был обучен на датасете, составленном из статей Википедии и литературных произведений [6]. Однако оказалось, что для работы в системах коммуникации с людьми такой набор обучающих данных не подходит, и его

требуется заменить на массив данных, состоящий из текстов, близких к разговорной речи людей. Сеть, обученная на постах из Интернета, получила название GPT-2. Увеличив количество параметров обучения более чем в 100 раз, разработчиками OpenAI была получена модель третьего поколения – GPT-3 [7].

Как и более ранние модели, GPT-3 оказалась наиболее эффективна в задачах генерации текста [6]. Предобученная GPT-3 – способ нагенерировать тексты на конкретно заданную тему, придумывать стихи в определенном стиле и давать ответы на вопросы по исходному тексту [7].

5. Алгоритм BERT

BERT (Bidirectional Encoder Representations from Transformers) представляет собой модель-трансформер с двунаправленным алгоритмом обучения, которое позволяет определять контекст слов как в прямом, так и в обратном порядке [8]. Двунаправленное обучение является отличительной чертой данной языковой модели от ее предшественников (рис. 3). Этот алгоритм позволяет модели BERT эффективно решать такие задачи, как:

– прогнозирование замаскированных слов (Masked LM);

– определять логическую связь между двумя отельными предложениями (Next Sentence Prediction – NSP).

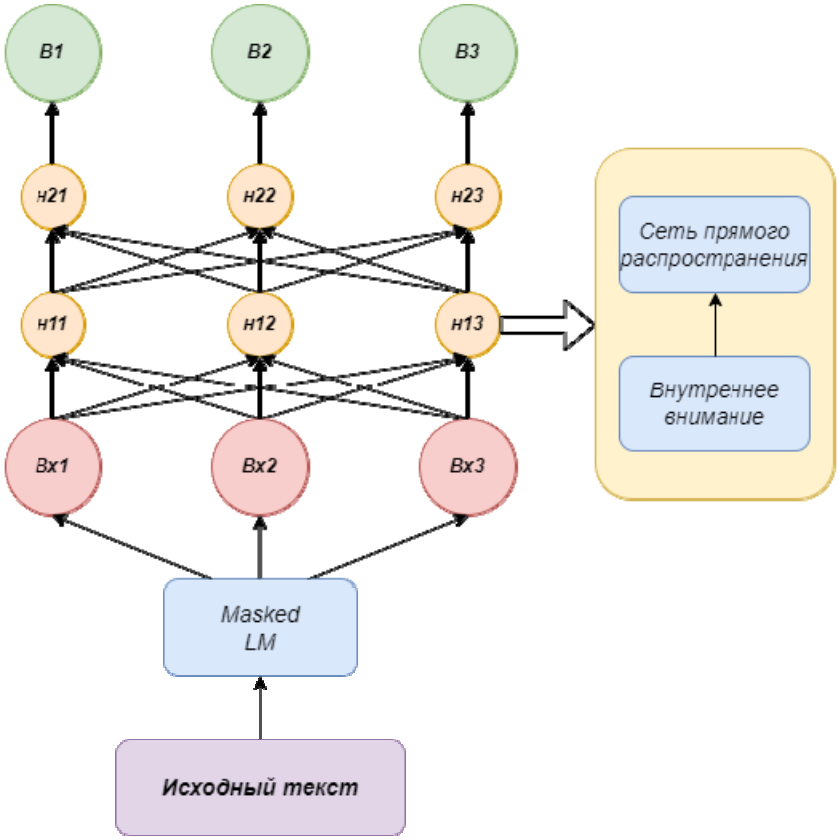

Рис. 3. Схема алгоритма BERT

Fig. 3. Scheme of the BERT algorithm

Суть прогнозирования замаскированных слов заключается в обучении модели правильно предугадывать стоящие рядом токены по их контексту и, наоборот, определять контекст каждого отдельного слова [9]. Таким образом, сеть определяет семантическую нагрузку каждого токена.

Установление логических связей между предложениями позволяет определять, как предложения связаны друг с другом, должно ли одно предложение предшествовать

другому или, наоборот, следовать за ним. Так, для модели становится возможным определения смысловой нагрузки не только одного слова, но и всего текста [8].

6. Алгоритм XLNet

XLNet (рис. 4) представляет собой языковую модель с архитектурой Transformer-XL. Отличительной чертой Transformer-XL от классической архитектуры Transformer является отсутствие фиксированной длины текста в условиях языкового моделирования без нарушения временной согласованности [10].

Таким образом, модель XLNet осуществляет обучение с помощью фиксированного прямого и/или обратного порядка факторизации, применяется Permutation Language Modeling – PLM, что в переводе означает «моделирование языка перестановок» [11]. Данный алгоритм объединяет идею авторегрессионного (характерного для GPT) и двунаправленного контекстного (характерного для BERT) моделирования.

Алгоритм PLM предполагает максимизацию ожидаемой логарифмической вероятности по всем возможным перестановкам последовательности, то есть данный алгоритм обучает авторегрессионную модель с различными перестановками токенов в предложении без маскировки отдельных слов (Masked LM). Благодаря этому контекст для каждого токена может состоять из лексем как слева, так и справа. Таким образом, вычисляется контекстуальная информация из всех позиций слов в предложении [11].

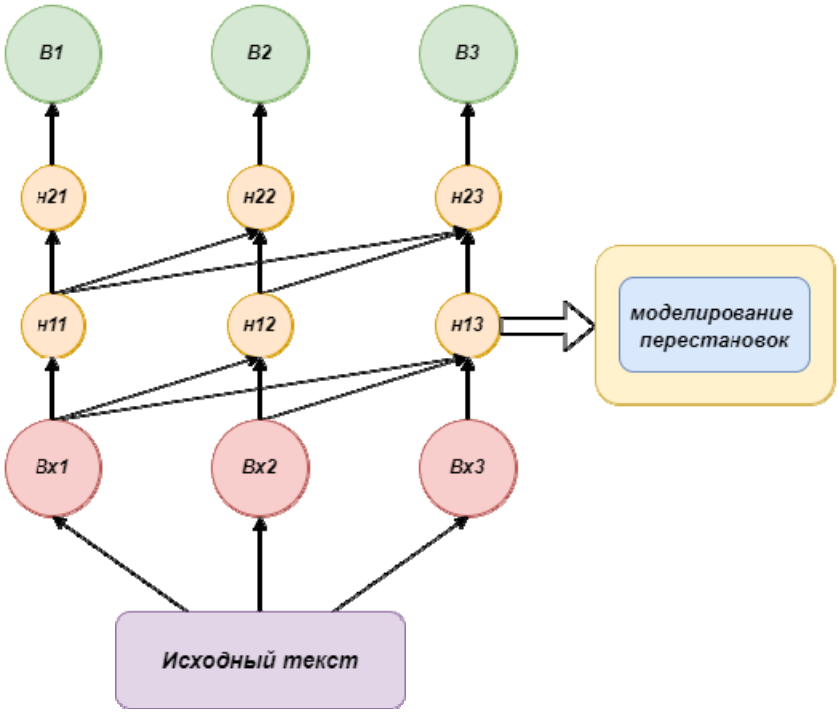

Рис. 4.Схема алгоритма XLNet

Fig. 4. Scheme of the XLNet algorithm

7. Анализ алгоритмов с архитектурой Трансформер с целью применения их для решения проблемы контроля качества

Модель GPT является однонаправленной и обладает авторегрессией. Данная языковая модель представляет собой алгоритм трансформер-декодера, подходящий для таких задач, как генерация текстов на различные темы [6]. Модель BERT, получившая большую известность, имеет противоположный алгоритм работы. Данная модель представляет собой трансформер-энкодер, не обладает регрессией и является двунаправленной, что позволяет ей справляться с такими задачами, как семантический анализ данных [9].

В задачах, требующих возможности обучения на небольших выборках, более эффективным будет применение модели GPT за счет наличия алгоритмов авторегрессии. Модель BERT требует больших обучающих данных [9]. Таким образом, прежде чем внедрять данную языковую модель в реальную среду, необходимо обучить ее на данных предметной области. В то же время модель GPT не требует предварительного обучения и тонкой настройки [7].

Однако за счет отсутствия двунаправленного обучения модель GPT не способна определять двусторонний контекст токенов, так как данный алгоритм предполагает обучение только по предшествующим слову позициям.

Таким образом, алгоритм BERT больше подходит для обработки входной информации (анализ отзывов, распознавание речи), а модель GPT – для генерации ответов системы.

Основным отличием модели XLNet от рассмотренных выше алгоритмов является отсутствие контекстных ограничений в виде фиксированной длины в условиях языкового моделирования и обучения за счет маскировки токенов. Данный алгоритм так же, как и GPT, обладает авторегрессий. Тем не менее, в отличие от GPT, в XLNet решена проблема одностороннего контекста. Это стало возможным благодаря механизму Permutation Language Modeling, так как эта модель состоит из алгоритма повторений и перестановок на уровне сегмента и новой схемы позиционного кодирования. Таким образом, XLNet не только позволяет фиксировать долгосрочные зависимости, но и решает проблему фрагментации контекста.

В отличие от алгоритма BERT в XLNet отсутствуют механизмы маскировки токенов для дальнейшего обучения, что гарантирует сохранность семантической значимости исходных данных при обработке [11]. Отсутствие маскировки данных позволяет механизму Self-Attention захватывать весь контекст предложения для определения смысловой нагрузки всей последовательности данных, поступающих на вход. Другим недостатком BERT относительно XLNet является наличие скрытых токенов при настройке модели и их отсутствие при использовании уже предобученной модели [11].

Однако, несмотря на присутствующие недостатки в решении многих задач, алгоритм BERT показал высокие результаты, превосходящие результаты GPT, но обойти результаты, показанные алгоритмом XLNet, с помощью модели BERT все же не удалось [11]. В ходе исследований были получены данные экспериментов, представленные на рисунке 5.

RACE	Accuracy	Middle	High
GPT [28]	59.0	62.9	57.4
BERT [25]	72.0	76.6	70.1
BERT+DCMN* [38]	74.1	79.5	71.8
RoBERTa [21]	83.2	86.5	81.8
XLNet	**85.4**	**88.6**	**84.0**

Рис. 5.Сравнение результатов, показанных современными алгоритмами, базирующимися на архитектуре Трансформер

Fig. 5. Comparison of the results shown by modern algorithms basing on the Transformer architectur

Эксперимент, в ходе которого были получены данные результаты, заключался в определении понимания прочитанного по набору тестов RACE и ClueWeb09-B. “Middle” и “High” в RACE – это два подмножества, представляющие размеры модели.

Так, модель XLNet, обладающая авторегрессией и двунаправленным обучением за счет механизма PLM, может прекрасно справляться с такими задачами, как языковое моделирование.

Сравнительная характеристика рассмотренных алгоритмов анализа естественного языка, базирующихся на архитектуре Трансформер, представлена в таблице 1.

Таблица 1

Сравнение алгоритмов с архитектурой Трансформер

Table 1. Comparison of algorithms with Transformer architecture

	Авторегрессия	*Двунаправленное обучение*	*Обучение без маскировки данных*	*Обучение на малых выборках*
GPT	✓	✗	✓	✓
BERT	✗	✓	✗	✗
XLNet	✓	✓	✓	✓

Выводы

Проведенный сравнительный анализ алгоритмов-трансформеров показал, что каждый из рассмотренных методов имеет свои преимущества и недостатки, каждый из них имеет свою область конкретных решаемых задач, для которых встроенные в них механизмы предназначены оптимальным образом. Для построения сложных систем, требующих решения задач из различных областей, следует применять программные комплексы, включающие в себя сразу несколько различных алгоритмов.

Для создания системы контроля качества обслуживания в организации требуется собрать программный комплекс, способный распознавать устную и письменную речь, обрабатывать ее, производя семантический анализ и анализ тональности данных, а также генерировать наиболее подходящие способы решения задачи и ответы на запросы клиентов. Так, для задачи распознавания речи эффективным будет использование модели BERT, для непосредственной обработки данных – XLNet, а для генерации ответов и решений – алгоритм GPT.

Примечания

1. Каннер Д.Д. Эффективная система корпоративного управления на предприятии сферы услуг (на примере сферы event-менеджмента) // Инновации. Наука. Образование. 2021. № 36. С. 1986–1993.

2. Интеллектуальные информационные системы: учеб. пособие / А.П. Частиков, К.И. Костенко, И.Ю. Леднева [и др.]. Краснодар, 2005. 327 с.

3. Белов Д.Л., Антипова О.Ю., Частикова В.А. Методы решения задач с конфликтными ситуациями в системах принятия решений // Труды Кубанского государственного технологического университета. 2000. Т. 7, № 1. С. 153–159.

4. Solomin A.A., Ivanova Yu.A. Modern approaches to multiclass intent classification based on pre-trained transformers // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2020. Vol. 20, No 4. P. 532–538. DOI: 10.17586/2226-1494-2020-20-4-532-538

5. Attention Is All You. Need / Ashish Vaswani, Noam Shazeer, Niki Parmar [et al.] // arXiv:1706.03762v5. 2017. DOI: 10.48550/ARXIV.1706.03762

6. Тюрюмина В.А. Разработка нейросети GPT-3 на базе NLP // Современные достижения молодежной науки: сб. ст. Междунар. науч.-исслед. конкурса, Петрозаводск, 11 мая 2021 года. Петрозаводск: Международный центр научного партнерства «Новая Наука» (ИП Ивановская Ирина Игоревна), 2021. С. 14–18.

7. Бажин В.А. Тонкая настройка BERT и GPT-3 для решения задачи генерации русскоязычных новостей // Актуальные научные исследования в современном мире. 2021. № 5-2 (73). С. 43–58.

8. Частикова В.А., Гуляй В.Г. Применение методов обработки естественного языка для решения задач обнаружения атак социальной инженерии // XII международная научно-практическая конференция молодых ученых, посвященная 61-ой годовщине полета Ю.А. Гагарина в космос: сб. науч. ст., Краснодар, 12–13 апреля 2022 года. Краснодар: Издательский Дом – Юг, 2022. С. 261–264.

9. Частикова В.А., Гуляй В.Г. Методика обнаружения атак социальной инженерии на основе алгоритмов анализа естественного языка // Прикаспийский журнал: управление и высокие технологии. 2022. № 3 (59). С. 61–71. DOI: 10.54398/20741707_2022_3_61

10. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context / Zihang Dai, Zhilin Yang, Yiming Yang, [et al.] // arXiv:1901.02860v3. 2019. DOI: 10.48550/ARXIV.1901.02860

11. XLNet: Generalized Autoregressive Pretraining for Language Understanding / Zhilin Yang, Zihang Dai, Yiming Yang [et al.] // arXiv:1906.08237. 2020. DOI: 10.48550/ARXIV.1906.08237

References

1. Kanner D.D. An effective corporate governance system at a service sector enterprise (on the example of the event-management sphere) // Innovations. Science. Education. 2021. No. 36. P. 1986–1993.

2. Intelligent information systems: a manual / A.P. Chastikov, K.I. Kostenko, I.Yu. Ledneva [et al.]. Krasnodar, 2005. 327 p.

3. Belov D.L., Antipova O.Yu., Chastikova V.A. Methods for solving problems with conflict situations in decision-making systems // Proceedings of the Kuban State Technological University. 2000. Vol. 7, No. 1. P. 153–159.

4. Solomin A.A., Ivanova Yu.A. Modern approaches to multiclass intent classification based on pre-trained transformers // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2020. Vol. 20, No 4. P. 532–538. DOI: 10.17586/2226-1494-2020-20-4-532-538

5. Attention Is All You. Need / Ashish Vaswani, Noam Shazeer, Niki Parmar [et al.] // arXiv:1706.03762v5. 2017. DOI: 10.48550/ARXIV.1706.03762

6. Tyuryumina V.A. Development of the GPT-3 neural network based on NLP // Modern achievements of youth science: coll. of art. of International scientific research competition, Petrozavodsk, May 11, 2021. Petrozavodsk: International Center for Scientific Partnership “New Science” (IP Ivanovskaya Irina Igorevna), 2021. P. 14–18.

7. Bazhin V.A. Fine adjustment of BERT and GPT-3 to solve the problem of generating Russian news // Actual scientific research in the modern world. 2021. No. 5-2 (73). P. 43–58.

8. Chastikova V.A., Gulyay V.G. Application of natural language processing methods for solving problems of detecting social engineering attacks // XII International scientific and practical conference of young scientists dedicated to the 61st anniversary of Yu.A. Gagarin’s flight into space: coll. of scientific articles, Krasnodar, April 12–13, 2022. Krasnodar: Publishing House – Yug, 2022. P. 261–264.

9. Chastikova V.A., Gulyay V.G. Methodology of social engineering attack detection based on natural language analysis algorithms // Caspian Journal: Management and High Technologies. 2022. No. 3 (59). P. 61–71. DOI: 10.54398/20741707_2022_3_61

10. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context / Zihang Dai, Zhilin Yang, Yiming Yang, [et al.] // arXiv:1901.02860v3. 2019. DOI: 10.48550/ARXIV.1901.02860

11. XLNet: Generalized Autoregressive Pretraining for Language Understanding / Zhilin Yang, Zihang Dai, Yiming Yang [et al.] // arXiv:1906.08237. 2020. DOI: 10.48550/ARXIV.1906.08237

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflicts of interests.

Статья поступила в редакцию 29.11.2022; *одобрена после рецензирования 18.12.2022*; *принята к публикации 19.12.2022*.

The article was submitted 29.11.2022; *approved after reviewing 18.12.2022*; *accepted for publication 19.12.2022*.