**Инженерный вестник Дона**, №7 (2024)
ivdon.ru/ru/magazine/archive/n7y2024/9371

# Эволюция и современное состояние систем ответов на вопросы: технологии распознавания намерений и именованных сущностей с использованием модели BERT

*К.А. Аксенов, Л. Сунь*

*Уральский федеральный университет имени первого Президента России Б.Н. Ельцина*

**Аннотация:** В данной статье подробно исследуется технологическая эволюция и текущее состояние вопросно-ответных систем. На примере задачи обслуживания клиентов авиакомпании разработана модель на основе BERT-модели, способная распознавать намерения пользователей и извлекать именованные сущности. В работе предоставлено подробное описание подготовки набора данных, методов анализа данных и методов исследования данных в рамках проекта. Представлено описание модели и настроек параметров во время процесса настройки модели и процесса ее обучения. Разработанная в этом проекте модель названа IRERAIR-BERT, которая достигла точности распознавания намерений 98,2% и точности распознавания именованных сущностей 83%. Мы создали модуль семантического анализа для системы вопросов и ответов. Следующим этапом нашей работы будет интеграция набора данных для завершения компонентов “запрос-ответ” и “генерация ответа” системы вопросов и ответов.
**Ключевые слова:** Вопросно-ответные системы, машинное обучение, нейронные сети, предобученные модели, распознавание намерений, распознавание именованных сущностей, анализ данных, обучение модели, трансформеры, диалоги.

## Обзор технологии системы вопросно-ответных систем и текущего состояния разработки

Вопросно-ответные системы (ВОС) играют крайне важную роль в повседневной жизни, позволяя людям взаимодействовать с интеллектуальными устройствами благодаря технологиям обработки естественного языка. Эти системы широко используются в таких областях, как виртуальные ассистенты (помощники), онлайн-поддержка, медицинские консультации и образование, что значительно улучшает доступность и эффективность получения информации. Более того, вопросно-ответные системы играют ключевую роль в улучшении пользовательского опыта, снижении затрат на рабочую силу и содействии персонализированным услугам, что делает их важным инструментом для реализации интеллектуальных услуг и повышения качества принятия решений. ВОС

отличаются от традиционных систем информационного поиска тем, что они используют технологии обработки естественного языка (ОЕЯ) и экспертные системы (ЭС), позволяя понимать запросы пользователей на естественном языке и предоставлять соответствующие решения. Выделяют вопросно-ответные системы общего назначения, а также предметно-ориентированные (для широких и узких предметных областей соответственно). Предметно-ориентированная ВОС обучается на определенном наборе знаний или данных, связанных с определенной предметной областью, что позволяет им предоставлять точные и сфокусированные ответы в этой области, данный тип систем в основном используют аппарат экспертных систем. ВОС общего назначения разработаны для ответа на вопросы по широкому кругу тем без ограничения на конкретную область. Эти системы обычно используют методы обработки естественного языка [1] нейронные сети [2], машинное обучение [3], для понимания и генерации ответов на широкий спектр вопросов.

Основная архитектура ранних ВОС заключалась в реализации отображения из "естественного языка" в запросы к базе данных [4], все они зависели от шаблонов правил, созданных вручную, и предварительно размеченных баз данных для преобразования вопросов на естественном языке в структурированные запросы. Однако эти системы были ограничены своей базой знаний и не могли ответить на вопросы, выходящие за ее рамки. До 1990-х годов предметно-ориентированные экспертные системы вопросов и ответов были основным направлением развития в этой области. Эти системы обеспечивали высокую точность и стабильные ответы в определенных областях, но они были ограничены своей однократной функциональностью и не обладали гибкостью, что приводило к ограничениям в их работе.

С 2017 года применение технологии глубоких нейронных сетей в области обработки естественного языка сделало революционные достижения. Начиная с ранних рекуррентных нейронных сетей и их вариантов с долгосрочной памятью, через сверточные нейронные сети [5], сети с указателями, до революционной модели "Трансформер" [6] и ряда предварительно обученных языковых моделей на ее основе, последовательное появление этих технологий значительно ускорило развитие ВОС. В некоторых задачах производительность систем вопросов и ответов постепенно достигла и даже превзошла человеческий уровень. В этот период также произошло развитие баз данных, используемых в ВОС. Взрывной рост данных в социальных сетях, использующих идеологию вопросно-ответного обмена [7] предоставил огромное количество пар вопрос-ответ для вопросно-ответных систем для их последующего анализа. В зависимости от источника используемого набора данных в предметной области выделяют два основных типа систем вопросов и ответов: основанные на часто задаваемых вопросах и основанные на данных диалогов.

Большое развитие в нейронные сети в 2018 году внесла языковая модель, основанная на архитектуре трансформер (англ. Bidirectional Encoder Representations from Transformers (BERT)) [8] от Google. BERT использует инновационную модель маскированного языка, основанную на двунаправленной архитектуре трансформера.

В конце 2022 года компания OpenAI представила генеративную предобученную модель-трансформер (англ. Generative pre-trained transformer (GPT) [9]) - языковую модель, специализированную на генерации диалогов. GPT способен не только точно понимать намерение пользователя, но и участвовать в эффективных многосценарных беседах (когда в одном диалоге обсуждаются разные вопросы из разных предметных областей или решаются

задачи междисциплинарного характера), предоставляя всесторонние, фокусированные и логически связанные ответы.

Необходимо уточнить, что упомянутый обзор предназначен для помощи читателям в более глубоком рассмотрении развития и текущего состояния систем ответов на вопросы. Развитие методов и средств машинного обучения, нейронных сетей, предобученных моделей и технологий, таких как GPT, направлено на повышение интеллекта и точности этих систем. Хотя вопросно-ответные системы общего назначения демонстрируют широкие возможности предоставления информационных услуг, это не означает, что технологическая сложность специализированных ВОС уступает технологической сложности универсальных систем. Во втором разделе описаны результаты разработки модуля "Распознавание намерений и именованных сущностей", который является основным компонентом ВОС. С помощью этого модуля мы проанализируем и обсудим технические требования к созданию систем ответа на вопросы в специализированных предметных областях.

## Распознавание намерений и сущностей

Чрезвычайно важный компонент эффективной ВОС - точная идентификация намерений пользователя из его сообщений, а также точное извлечение именованных сущностей, имеющих отношение к этим намерениям [10, 11]. Модуль текстового анализа системы отвечает за распознавание намерений пользователя и извлечение именованных сущностей из предоставленной информации. Эта информация затем передается модулю запросов, который связывается с соответствующей базой данных на основе намерений пользователя. Запрос основан на именованных сущностях, выявленных модулем понимания текста, чтобы найти и сопоставить наиболее релевантные ответы. Система выводит наиболее вероятный правильный ответ пользователю. В настоящее время

представители отрасли предпочитают использовать настроенные большие языковые модели (БЯМ) [12] или генеративные БЯМ, такие как программный интерфейс приложения (англ. Application Programming Interface (API)), предоставляемый GPT [13], для улучшения производительности различных модулей в ВОС. Тем не менее, не все системы вопросов и ответов подходят для создания с использованием БЯМ, и решение должно быть основано на конкретных требованиях.

**Рискованность процесса моделирования**

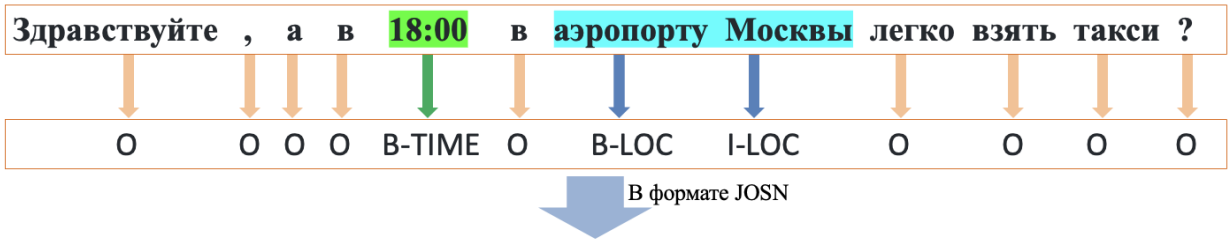
GPT может выполнять задачу распознавания именованных сущностей с использованием “Техники подсказок” (англ. prompt-based learning), но не позволяет получать качественные результаты, как показано в [14]. Для обеспечения стабильности в индустрии широко применяется метод тонкой настройки предварительно обученных моделей, таких как ERNIE и RoBERTa. Архитектура этих моделей обычно основана на фреймворке трансформер. Преимущество тонкой настройки предварительно обученных моделей заключается в использовании параметров или весов модели для достижения высокопроизводительного семантического понимания и стабильных возможностей распознавания признаков в конкретных профессиональных областях.

Как упоминалось ранее, в ВОС распознавание намерений пользователя связано с конкретными предметными областями, за которыми следует идентификация и подготовка ответов на основе соответствующей базы данных предметной области. Поэтому перед тонкой настройкой модели необходимо провести инжиниринг признаков на данных сущностей и намерений. Изначально объем данных должен составлять тысячи записей. Если проект требует множество типов меток, для надежной поддержки необходимо большое количество данных. Затем, в зависимости от требований проекта, данные необходимо аннотировать. Распространенные

типы намерений и именованных сущностей включают “IOB-разметку последовательностей”, “BIOES-разметку последовательностей” и “BIES-разметку последовательностей” [15, 16]. Среди них буквы «B», «I», «O», «E», «S» представляют следующие значения: начало (Begin-B), внутри (Inside-I), снаружи (Outside-O), конец (End-E), единственный (Single-S).

На рис.1 приведен пример набора данных для распознавания намерений и идентификации именованных сущностей. После аннотации данных важно оценить баланс распределения меток в наборе данных. Если имеется значительный дисбаланс в распределении меток, требуются корректировки в наборе данных.

Обычно задачи аннотации могут быть упрощены с использованием визуальных платформ аннотации или библиотек на Python, таких как “SpaCy”, “NLTK” и “CoNLP”, для выполнения процесса аннотации. Распределение меток может быть изучено с использованием библиотеки pandas для анализа данных.




```
{ "text": "Здравствуйте, а в 18:00 в аэропорту Москвы легко взять такси?",
"intent" : " atis_ground_service"
"entities": [
        {"entity": "Здравствуйте", "tag": "O"},
        {"entity": ",", "tag": "O"},
        {"entity": "а", "tag": "O"},
        {"entity": "в", "tag": "O"},
        {"entity": "18:00", "tag": "B-TIME"},
        {"entity": "в", "tag": "O"},
        {"entity": "аэропорту", "tag": "B-LOC"},
        {"entity": "Москвы", "tag": "I-LOC"},
        {"entity": "легко", "tag": "O"},
        {"entity": "взять", "tag": "O"},
        {"entity": "такси", "tag": "O"},
        {"entity": "?", "tag": "O"}
        ]
}
```

Рис. 1. – Метод разметки IOB и формат Json.

Для решения практических задачи распознавания сущностей и намерений обслуживания клиентов авиакомпании (англ. Intent recognition and entity recognition in air services (IRERAIR) нами была разработана на основе модели BERT модель IRERAIR-BERT, для которой выделены следующие типы тегов: 8 типов меток намерений и 122 типа меток именованных сущностей, как показано на Рис. 2 и Рис. 3. В таблице № 1 представлен обзор набора данных для данного проекта.

```
[ ] label2id = {k: v for v, k in enumerate(final_df.tag.unique())}
    id2label = {v: k for v, k in enumerate(final_df.tag.unique())}
    label2id.head()

{'O': 0,
 'B-fromloc.city_name': 1,
 'B-depart_time.time': 2,
 'I-depart_time.time': 3,
 'B-toloc.city_name': 4,
 'B-arrive_time.time': 5,
 'B-arrive_time.period_of_day': 6,
 'B-depart_date.day_name': 7,
 'B-depart_time.period_of_day': 8,
 'B-flight_time': 9,
 'I-flight_time': 10,
 'I-fromloc.city_name': 11,
 ............
'B-return_date.today_relative': 120,
'I-return_date.today_relative': 121,
'B-return_date.day_name': 122}
```

Рис. 2. – Фрагменты разделов 122 меток именованных сущностей модели IRERAIR-BERT

```
[ ] for key,value in enumerate(labels):
      print(value)

atis_flight
atis_quantity
atis_flight_time
atis_abbreviation
atis_airfare
atis_ground_service
atis_airline
atis_aircraft
```

Рис. 3. – Восемь этикеток намерений для модели IRERAIR-BERT

Таблица № 1

Набор данных по обслуживанию клиентов авиакомпаний

| Общий объем данных, текстов запросов | Типов меток намерений | Типов меток сущностей |
|---|---|---|
| 1480 | 8 | 122 |

После подготовки данных необходимо выбрать подходящую модель, обычно учитывая такие факторы, как количество параметров, условия аппаратного обеспечения для выполнения и разнообразие поддерживаемых языков. В данном проекте выбрали мультиязычную модель BERT-BASE с 109 миллионами параметров. Данные были разделены на тренировочный, тестовый и валидационный наборы в соотношении 2:1:1, соответственно. После этого провели настройку параметров обучения со следующими настройками: Эпохи = 3, Размер пакета = 16 и Скорость обучения = 0.0001. На рис. 4 показана модель, созданная с использованием метода настройки BERT-модели. Таким образом разработанная модель IRERAIR-BERT способна распознавать намерения пользователей и извлекать именованные сущности, необходимые для ВОС авиакомпании. Результаты тестирования, представленные в Таблице № 2, показывают точность распознавания намерений 98.2% и точность распознавания именованных сущностей 83%.

Таблица № 2

Точность распознавания намерений и распознавания именованных сущностей и оценка полноты (англ. Harmonic mean of precision and recall. (F1)) для модели IRERAIR-BERT

| Тип задания | Модель | Точность | Оценка F1 |
|---|---|---|---|
| Намерение | IRERAIR-BERT | 0.985 | 0.941 |
| Сущность | IRERAIR-BERT | 0.83 | 0.80 |

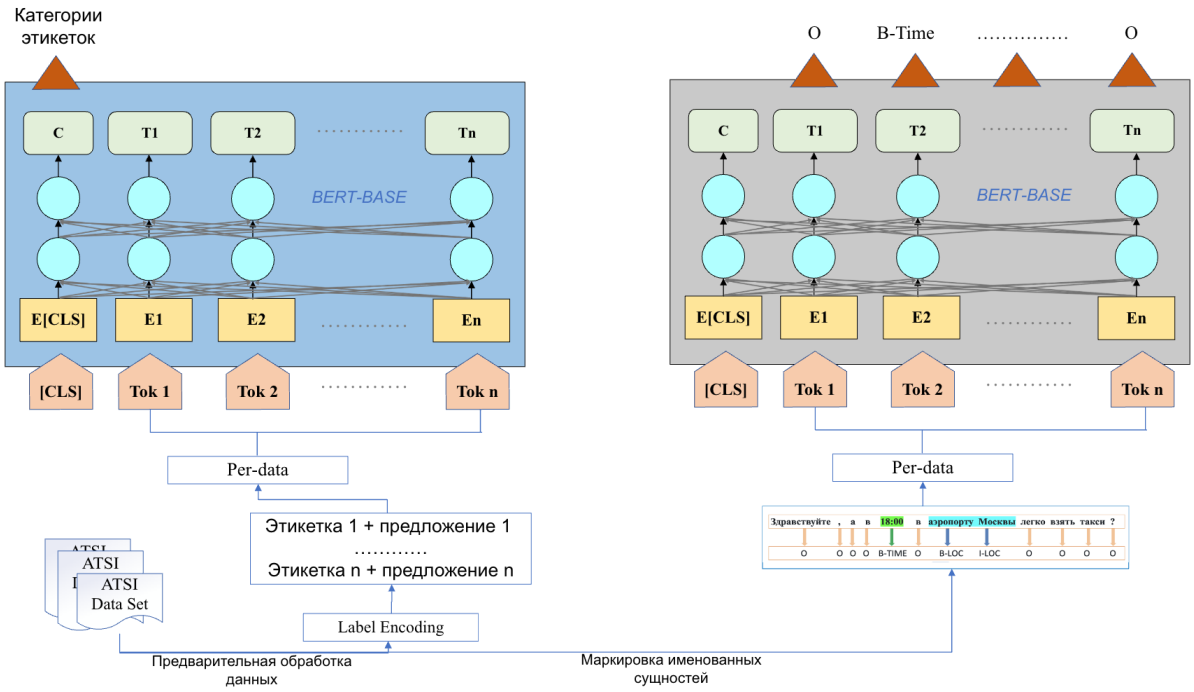



Рис. 4. – Схема процесса тонкой настройки модели.

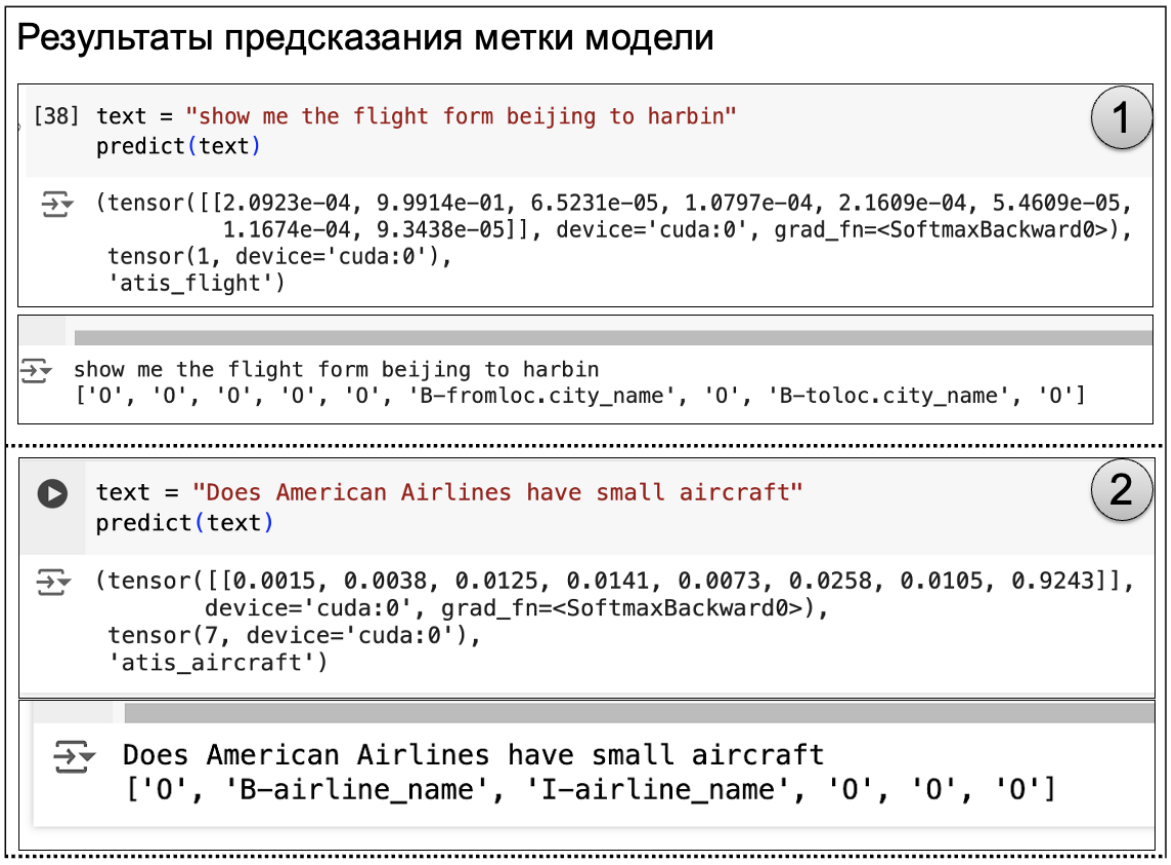



Рис. 5. – Результаты оценки тестов на распознавание намерений и именованных сущностей для модели IRERAIR-BERT

Результаты тестирования модели IRERAIR-BERT показали, что она автоматически распознаёт намерения, и точно идентифицирует именованные сущности. На рис.5 представлены входные данные и результаты диагностики.

## Заключение

В данной работе мы рассмотрели текущее состояние развития систем вопросно-ответных систем. На основе BERT-модели была разработана и настроена модель IRERAIR-BERT для задачи извлечения сущностей и намерений на этапе анализа запроса пользователя для процесса обслуживания клиентов авиакомпании. Применение модели демонстрирует точность распознавания намерений 98.2% и точность распознавания именованных сущностей 83%. В дальнейшем планируется решение задачи генерации ответа на пользовательский запрос.

## Литература

1. Arora R., Singh P., Goyal H., Vijayvargiya S. Comparative question answering system based on natural language processing and machine learning // IEEE International Conference on Artificial Intelligence and Smart Systems. 2021. Pp. 373-378.

2. Gemirter C.B., Goularas D.A. Turkish question answering system based on deep learning neural networks. Journal of Intelligent Systems: Theory and Applications. 2021. № 4. Pp.65-75.

3. Bian Y., Peng K. Question Answering System Analysis Based on Machine Learning // IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology. 2021. Pp.279-283.

4. Turing A.M. Computing Machinery and Intelligence. Parsing the Turing Test. 2009. Pp. 23-65.

5. Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015. № 61. Pp.85-117.

6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Lukasz K., Polosukhin I. Attention is all you need // Advances in neural information processing systems. 2017. № 30. Pp.5999-6009.

7. Srba I., Bielikova M. A comprehensive survey and classification of approaches for community question answering. ACM Transactions on the Web. 2016. № 10. Pp.1-63.

8. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. № 19-1423(1). Pp. 4171-4186.

9. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training. 2018.

10. Lihong, Xingchi, Yanhua, Malu, Liaoqun, Zhian, Liuliang, Lichao, Yunsen, Yongchao. Exploration and Practice of NER Technology in Meituan Search. URL: tech.meituan.com/2020/07/23/ner-in-meituan-nlp.html.

11. Харламов А.А., Ермоленко Т.В., Дорохина Г.В. Сравнительный анализ организации систем синтаксических парсеров. Инженерный вестник Дона, 2013. № 4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015.

12. Cavalin P., Ribeiro V.H.A., Appel A., Pinhanezet C. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. № 2020.emnlp-main.324. Pp.3952-3961.

13. OpenAI team. Gpt-4 technical report. 2023. Pp. 1-100.

14. Ray P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023. № 3. Pp.121-154.

15. Alshammari N., Alanazi S. The impact of using different annotation schemes on named entity recognition. Egyptian Informatics Journal. 2021, № 22(3). Pp.295-302.

16. Строцев В.А. Информативность частотных характеристик N-грамм текстовых фрагментов. Инженерный вестник Дона. 2013. № 1. URL: ivdon.ru/ru/magazine/archive/n1y2013/1492.

## References

1. Arora R., Singh P., Goyal H. IEEE International Conference on Artificial Intelligence and Smart Systems. 2021. Pp. 373-378.

2. Gemirter C.B., Goularas D.A. Journal of Intelligent Systems: Theory and Applications. 2021. № 4. Pp.65-75.

3. Bian Y., Peng K. IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology. 2021. Pp.279-283.

4. Turing A.M. Parsing the Turing Test. 2009. Pp. 23-65.

5. Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015. № 61. Pp.85-117.

6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J. Advances in neural information processing systems. 2017. № 30. Pp.5999-6009

7. Srba I., Bielikova M. ACM Transactions on the Web. 2016. № 10. Pp.1-63.

8. Devlin J, Chang M W, Lee K, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. № N19-1423(1). Pp. 4171-4186

9. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving language understanding by generative pre-training. 2018

10. Lihong, Xingchi, Yanhua, Malu, Liaoqun, Zhian, Liuliang, Lichao, Yunsen, Yongchao. Exploration and Practice of NER Technology in Meituan Search. URL: tech.meituan.com/2020/07/23/ner-in-meituan-nlp.html.

11. Xarlamov A.A., Ermolenko T.V., Doroxina G.V. Inzhenernyj vestnik Dona, 2013. № 4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015.

12. Cavalin P., Ribeiro V.H.A., Appel A. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. № 2020.emnlp-main.324. Pp.3952-3961.

13. OpenAI team. Gpt-4 technical report. 2023. Pp. 1-100

14. Ray P.P. Internet of Things and Cyber-Physical Systems. 2023. № 3. Pp.121-154.

15. Alshammari N., Alanazi S. Egyptian Informatics Journal. 2021, № 22(3). Pp.295-302.

16. Strocev V.A. Inzhenernyj vestnik Dona. 2013. № 1. URL: ivdon.ru/ru/magazine/archive/n1y2013/1492.

**Дата поступления: 19.05.2024**

**Дата публикации: 2.07.2024**