NLP and LLM Based Approach to Enterprise Knowledge Base Construction

Valery Taberko and Dzmitry Ivaniuk
JSC «Savushkin Product»
Brest, Belarus
Email: tab@pda.savushkin.by, id@pda.savushkin.by

Viktor Smorodin and Vladislav Prokhorenko
Department of Mathematical Problems
of Control and Informatics
Francisk Skorina Gomel State University
Gomel, Belarus
Email: smorodin@gsu.by, shnysct@proton.me

***Abstract*—An approach to the automated formation of a knowledge base about an enterprise is proposed based on the analysis of existing technical documentation using large language models. This approach makes it possible to integrate the proposed solution with other developments and enterprise software to ensure the construction of intelligent automated control systems, recommendation systems and decision support systems, and information support systems for enterprise personnel.**

***Keywords*—Industry 4.0, standard, large language model, NLP, knowledge extraction**

I. Introduction

Effective implementation of methods for automating control and decision making requires ensuring semantic compatibility of heterogeneous components of intelligent systems [1]. The OSTIS technology uses a multi-agent approach to the development of intelligent systems, when agents operate on a single knowledge base.

Thus, in the course of building intelligent systems that provide solutions to a class of problems of automated management and information support of an enterprise, the task arises of combining all information about the enterprise into a single information space - a knowledge base, which is stored in the semantic memory of the intelligent system. The sources of such knowledge can be existing descriptions of the work of enterprises within the framework of accepted international standards.

Information extraction (IE) is a fundamental challenge in the field of natural language processing (NLP), especially in the context of the modern digital era, where data volumes are rapidly increasing, as is the need for rapid processing [2]. The importance of IE is evident in a variety of applications, ranging from automating text analysis and extracting structured data from various sources to supporting decision making based on large volumes of information. Based on the principles of machine learning and deep learning, modern IE methods provide the ability to automatically extract information from texts and documents, which is an important step in the digital transformation of organizations and society as a whole.

In the framework of Industry 4.0, where digitalization and automation play a key role in industrial transformation, the importance of automating the extraction of information from technical documents becomes critical to building a digital twin of the enterprise. A digital twin of an enterprise is a virtual model of its physical processes, equipment and resources based on real-time data. Automated information extraction systems allow you to quickly and efficiently process technical documents, highlighting key parameters, equipment characteristics, production process structure and other important data necessary to create and maintain a digital twin.

Using a digital twin of an enterprise provides organizations with the opportunity to conduct virtual modeling and analysis of production processes, optimize their efficiency, predict possible problems and take preventive measures. This helps increase flexibility, respond to changes in real time and ensure more efficient use of enterprise resources. Thus, automation of information extraction from technical documents is an important link in the process of creating and maintaining a digital twin of an enterprise, which ultimately helps to increase its competitiveness and sustainability in the market.

Information support for employees of an industrial enterprise also plays a key role in ensuring competitiveness and production efficiency. With the development of digital technologies and the penetration of the Internet of things into production processes, employees are faced with a huge amount of data that requires analysis and decision-making. Information support not only allows you to effectively manage data, but also provides access to up-to-date information in real time, which is necessary for a quick response to changes in the production environment.

At the present stage of development of complex technological systems, the volume of information contained in technical documentation describing the relevant subject areas can be large and also difficult to analyze due to the heterogeneity and ambiguity of the interpretation of some provisions of the standards [3].

This paper proposes an approach to solving this problem: automation of knowledge extraction from documentation based on the use of modern neural network technologies - large language models.

II. Problems of working with standards when designing new generation intelligent systems

Existing international standards for describing production systems make it possible to release an ontological approach to solving problems of production automation, building recommender systems and enterprise information support [4] . Thus, we can highlight the correspondence of the concepts of the ontology of the subject area "probabilistic technological production processes" to the concepts from the ISA 5.1, ISA-88 and ISA-95 [1], [3], [5]–[7] standards.

At the same time, a number of problems arising in connection with the use of standards can be identified. One of the key difficulties is the heterogeneity and fragmentation of standards in various areas, which makes it difficult for them to interact and integrate into the overall system. These standards are often developed by different organizations and committees, resulting in a variety of formats, syntax and semantics, which makes them difficult to understand and interoperate within a single system. This creates compatibility problems between various system components, and also increases the complexity of the development and support process [8]–[10].

Another major challenge is the rapid development of technology, which leads to constant updating and modification of existing standards. With the development of new technologies, standards also evolve, but on the other hand, periodic changes and updates may lead to incompatibility between different versions of standards and technologies. This complicates the integration and support of intelligent systems and can also cause compatibility issues between different components and devices, making it difficult to communicate and communicate between them. Thus, successful implementation and maintenance of intelligent systems requires continuous updating and adaptation to changing standards and technologies.

Overcoming these difficulties is critical to the development of effective intelligent systems that can analyze, predict and optimize various aspects of enterprise functioning. This includes not only managing production processes, but also optimizing logistics schemes, forecasting demand, analyzing market trends and many other aspects of business. In this context, the use of modern methods of data processing, machine learning and artificial intelligence becomes a necessity to create adaptive and intelligent systems that can effectively respond to dynamically changing market conditions and the competitive environment.

III. Large language models

The historical path of natural language processing (NLP) before the advent of large language models is characterized by the gradual development of methods and approaches to text analysis. Since early research in linguistics and artificial intelligence, such as the creation of the first grammars and automatic translation systems, progress in NLP has been associated with finding ways for computers to efficiently understand and generate natural language. In the early periods of NLP development, the emphasis was on rules and symbolic approaches to text analysis, which often limited its applicability due to the difficulty of creating complete and accurate rules for different languages and contexts.

However, the true breakthrough came with the development of machine learning methods and deep learning in particular. The emergence of large language models in recent years has opened up wide opportunities for working with text information in various areas of human activity. These models, such as GPT [11] and BERT [12], have a unique ability to understand the semantics of natural language and can efficiently analyze and process large volumes of text data.

Research shows that the use of such models leads to significant improvements in natural language processing tasks, including machine translation, sentiment analysis, information extraction, and more [13]. Thanks to their capabilities, working with text information becomes more efficient and accurate, which opens up new prospects for the development of automated data processing systems and artificial intelligence.

In the context of the problem under consideration, it is important, in particular, to note significant progress in solving the problem of information extraction [14]. Integrating large language models into the process of extracting information from documents according to a given ontology is an important and promising direction of research in the field of natural language processing (NLP). These models have a unique ability to understand the semantics of the text and can effectively highlight key entities and relationships between them

At the moment, we can note existing proposals for integrating LLM with OSTIS Technology [15], [16], which use third-party services (ChatGPT [17])

It is also necessary to emphasize that despite the modern successes of using LLM in various fields of activity, it is also necessary to note their disadvantages, such as inconsistency and the possibility of errors. Such models can be highly sensitive to noise and anomalies in the input data, which can lead to unpredictable results and potential errors in text interpretation [18]. There is a well-known phenomenon of hallucination in large language models, which is a phenomenon in which the model generates content that may be unpredictable, unrelated to the context, or even fictitious [19]. This phenomenon is a consequence of the huge amount of data on which models are trained, as well as their complex architecture, which includes billions of parameters. Under certain conditions, models can produce text that appears consistent and plausible, but is actually a random combination of words and phrases [20].

Moreover, the interpretability of large language models poses a significant challenge. Due to their utilization of intricate deep learning architectures, the internal workings of these models often remain opaque and convoluted, hindering human comprehension. This lack of transparency raises concerns regarding the justification of the model's decisions and undermines the confidence in relying on its conclusions for critical decision-making processes. Without a clear understanding of how the model arrives at its outputs, stakeholders may hesitate to fully trust its recommendations or insights, thereby impeding the integration of these advanced AI systems into real-world applications where trust and accountability are paramount.

This phenomena highlight the need to critically analyze the results obtained from large language models and emphasizes the importance of developing methods to control the quality and reliability of their use.

IV. An approach to automated construction of a knowledge base about an enterprise from a standardized description

The main idea of the proposed approach is to implement a component of the OSTIS ecosystem that implements functionality for automated extraction of information from standardized technical documentation of an enterprise.

The importance of prompt engineering for large language models is emphasized in the context of ensuring accurate and efficient interaction with the model. Prompts, being the key interface between the user and the model, determine the formulation of the problem and the context of the request, which directly affects the quality and relevance of the answers. Prompt engineering allows you to structure the input data for the model, taking into account the specifics of the problem and the requirements of the application, which provides more accurate and relevant results [21]. Automatic generation of prompts for large language models is a key area of research aimed at improving their performance and adaptability to a variety of tasks and contexts.

To successfully solve the problem of extracting information from standardized documents in accordance with a given domain ontology, it is necessary to generate a prompt that will explicitly define the required information and provide the model with an understanding of the context of the document and the key concepts of probabilistic technological processes.

In the prompt, the target information should be explicitly defined, namely, a set of concepts for work, and the connections between them. In addition, it is necessary to take into account the specific structure and format of documents of a particular standard. Including examples of the target input and output in the prompt provides not only additional context for the model, but also a more explicit representation of the desired output. This approach allows the model to better capture user intent, recognize key query elements, and correlate them with corresponding output. In this way, the model becomes capable of generating more accurate and appropriate responses or actions, which significantly improves its functionality and meets the users' needs for better service.

Thus, during the operation of the automated knowledge extraction component, it is necessary to solve the following tasks:

- determine the standard of the document and carry out its preliminary preparation;
- generate a prompt containing a description of a specific standard and domain ontology, as well as example of the expected output;
- send a prompt to a large language model;
- receive the model's response and verify its correctness and compliance with the ontology;
- use the extracted knowledge from the answer in the knowledge base.

The general scheme of prompt formation is shown in Fig. 1.

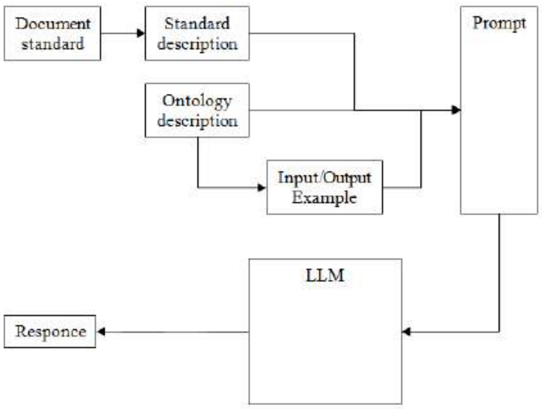


Figure 1. Scheme for generating a prompt for LLM

Within the framework of the OSTIS Technology, a multi-agent approach to problem solving [1] is used. This approach involves defining a hierarchy of agents that perform atomic functions for preparing and converting information. The proposed composite solver within this approach has the structure presented in Figure 2. Each agent specializes in certain aspects of information processing.

abstract non-atomic sc-agent of automated knowledge extraction from technical documentation

⇒ *decomposition of abstract sc-agent**:

{
- *abstract sc-agent of document preparation*
- *abstract sc-agent of standard description*
- *abstract sc-agent of ontology description*
- *abstract sc-agent of LLM communication*
- *abstract sc-agent of LLM responce parsing*

}

Figure 2. Decomposition of the abstract non-atomic sc-agent of automated knowledge extraction

The sc-agent of document preparation is responsible for the preliminary preparation of the document and determining the type of standard to which it belongs. The document is converted into text form. This process may be accompanied by the use of image analysis tools (if images are present in the document).

The sc-agent of standard description is capable of returning a description of the standard to which the document in question was defined. It conveys information about the specific organization of text in documents of a particular standard.

The sc-agent of ontology description agent is capable of returning a description of the ontology of the subject area (for example, “probabilistic technological production processes”): a set of concepts, their description and a description of the connections between them.

The sc-agent of LLM communication is responsible for generating a prompt from information collected from previous agents and sending it as a request to the LLM. After sending, the agent expects to receive a response. The agent is not tied to a specific implementation or location of a large language model.

The sc-agent of LLM response parsing is the key element responsible for checking the correctness of the response received from the large language model (LLM). This agent checks the LLM response for compliance with the data format and consistency with the domain ontology. Its functionality is necessary to account for possible situations of "hallucination" or errors that may occur in the operation of the knowledge extractor-LLM. If the answer is successfully verified, the extracted information can be used to populate the knowledge base. Given the complex nature of the task that this agent needs to solve it may operate also using an LLM.

V. An example of extracting information from documentation using LLM

An example of the information extraction approach can be demonstrated using the project documentation for the PLCnext testbed [22] (Fig. 3, 4, 5).

The documentation contains extensive text on software settings for working with the stand (Fig. 6). Using the described prompt engineering, you can use LLM (ChatGPT 3.5 [17]) to structured knowledge extraction for each configuration step in JSON format, where each instruction step is correctly separated into a separate element (Fig. 7 and 8).

Figure 3. General view of the stand

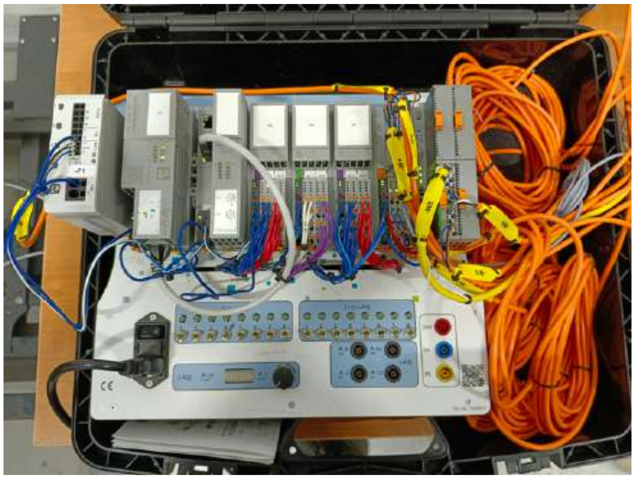


Figure 4. Controller and input/output node of the stand

VI. Conclusion

This paper proposes an approach to solving an important practical problem - automated knowledge extraction for constructing enterprise knowledge bases, based on the application of an agent-based approach within the framework of OSTIS Technology and the use of large language models. The logic of operation of the corresponding component of the OSTIS Ecosystem is presented, which allows you to systematize and retrieve information from various data sources. The described approach can be applied not only in the field of enterprises, but also

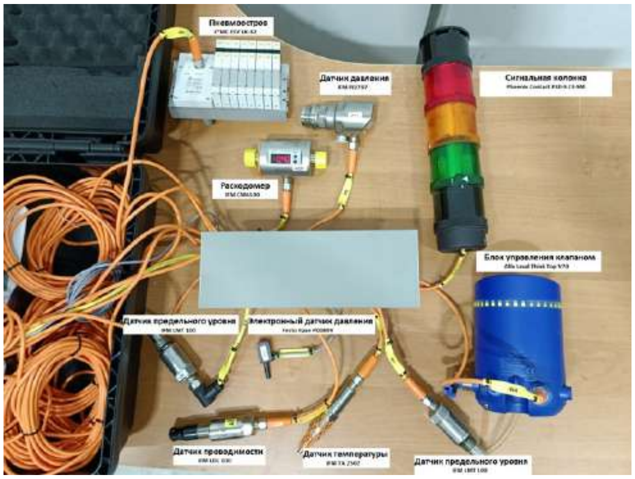
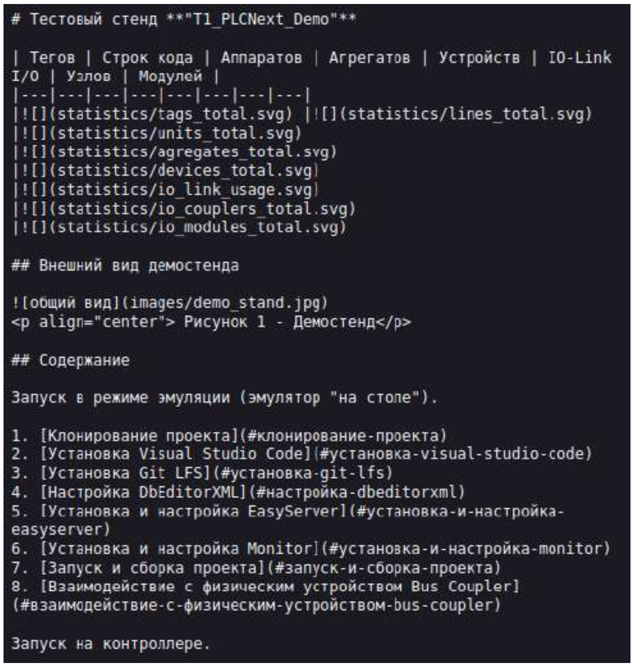
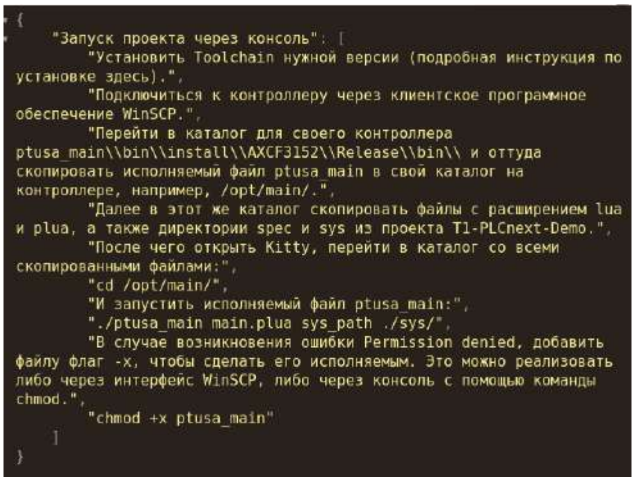


Figure 5. Sensors and actuators



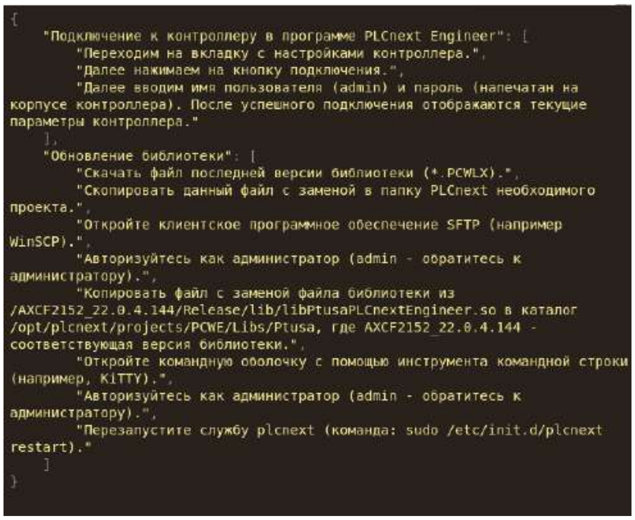
```
# Тестовый стенд **"T1_PLCNext_Demo"**

| Тегов | Строк кода | Аппаратов | Агрегатов | Устройств | IO-Link
I/O | Узлов | Модулей |
|---|---|---|---|---|---|---|---|
|![](statistics/tags_total.svg) |![](statistics/lines_total.svg)
|![](statistics/units_total.svg)
|![](statistics/agregates_total.svg)
|![](statistics/devices_total.svg)
|![](statistics/io_link_usage.svg)
|![](statistics/io_couplers_total.svg)
|![](statistics/io_modules_total.svg)

## Внешний вид демостенда

![общий вид](images/demo_stand.jpg)
<p align="center"> Рисунок 1 - Демостенд</p>

## Содержание

Запуск в режиме эмуляции (эмулятор "на столе").

1. [Клонирование проекта](#клонирование-проекта)
2. [Установка Visual Studio Code](#установка-visual-studio-code)
3. [Установка Git LFS](#установка-git-lfs)
4. [Настройка DbEditorXML](#настройка-dbeditorxml)
5. [Установка и настройка EasyServer](#установка-и-настройка-
easyserver)
6. [Установка и настройка Monitor](#установка-и-настройка-monitor)
7. [Запуск и сборка проекта](#запуск-и-сборка-проекта)
8. [Взаимодействие с физическим устройством Bus Coupler]
(#взаимодействие-с-физическим-устройством-bus-coupler)

Запуск на контроллере.
```

Figure 6. Original view of the document text



```
{
    "Запуск проекта через консоль": [
        "Установить Toolchain нужной версии (подробная инструкция по
установке здесь).",
        "Подключиться к контроллеру через клиентское программное
обеспечение WinSCP.",
        "Перейти в каталог для своего контроллера
ptusa_main\\bin\\install\\AXCF3152\\Release\\bin\\ и оттуда
скопировать исполняемый файл ptusa_main в свой каталог на
контроллере, например, /opt/main/.",
        "Далее в этот же каталог скопировать файлы с расширением lua
и plua, а также директории spec и sys из проекта T1-PLCnext-Demo.",
        "После чего открыть Kitty, перейти в каталог со всеми
скопированными файлами:",
        "cd /opt/main/",
        "И запустить исполняемый файл ptusa_main:",
        "./ptusa_main main.plua sys_path ./sys/",
        "В случае возникновения ошибки Permission denied, добавить
файлу флаг -x, чтобы сделать его исполняемым. Это можно реализовать
либо через интерфейс WinSCP, либо через консоль с помощью команды
chmod.",
        "chmod +x ptusa_main"
    ]
}
```

Figure 7. Structured Information Retrieval



```
{
    "Подключение к контроллеру в программе PLCnext Engineer": [
        "Переходим на вкладку с настройками контроллера.",
        "Далее нажимаем на кнопку подключения.",
        "Далее вводим имя пользователя (admin) и пароль (напечатан на
корпусе контроллера). После успешного подключения отображаются текущие
параметры контроллера."
    ],
    "Обновление библиотеки": [
        "Скачать файл последней версии библиотеки (*.PCWLX).",
        "Скопировать данный файл с заменой в папку PLCnext необходимого
проекта.",
        "Откройте клиентское программное обеспечение SFTP (например
WinSCP).",
        "Авторизуйтесь как администратор (admin - обратитесь к
администратору).",
        "Копировать файл с заменой файла библиотеки из
/AXCF2152_22.0.4.144/Release/lib/libPtusaPLCnextEngineer.so в каталог
/opt/plcnext/projects/PCWE/Libs/Ptusa, где AXCF2152_22.0.4.144 -
соответствующая версия библиотеки.",
        "Откройте командную оболочку с помощью инструмента командной строки
(например, KiTTY).",
        "Авторизуйтесь как администратор (admin - обратитесь к
администратору).",
        "Перезапустите службу plcnext (команда: sudo /etc/init.d/plcnext
restart)."
    ]
}
```

Figure 8. Structured Information Retrieval

in other subject areas where work with standardized documentation is required as a source for populating knowledge bases.

References

[1] V. Golenkov, Ed., Tehnologija kompleksnoj podderzhki zhiznennogo cikla semanticheski sovmestimyh intellektual'nyh komp'juternyh sistem novogo pokolenija [Technology of complex life cycle support of semantically compatible intelligent computer systems of new generation]. Bestprint, 2023.

[2] C. D. Manning, H.Schütze Foundations of statistical natural language processing, 1999, MIT press.

[3] V. Taberko, D. Ivaniuk, V. Smorodin, V. Prokhorenko Adaptive Control System for Technological Process within OSTIS Ecosystem, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 291-298.

[4] V. Taberko, D. Ivaniuk, N. Zotov, M. Orlov, O. Pupena, N. Lutska Principles Of Building A System For Automating The Activities Of A Process Engineer Based On An Ontological Approach Within The Framework Of The Industry 4.0 Concept, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2021, pp. 209-218.

[5] "ISA5.1 Standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa5-1/, (accessed 2024, Jan)

[6] "ISA-88 standard," Available at: https://www.isa.org/isa88/, (accessed 2024, Jan).

[7] "ISA-95 standard," https://www.isa.org/standards-and-publications/isa-standards/isa-standards-committees/isa95/, (accessed 2024, Jan)

[8] J. Lee, B. Choi A survey of standardization efforts for service-oriented architecture, Expert Systems with Applications, 2009, 36(4), pp.7844-7855.

[9] M. Salama, S.K. Mostefaoui, A. Nour A survey on IoT interoperability and standardisation issues, International Journal of Ad Hoc and Ubiquitous Computing, 13(3), 145-155.

[10] V. Taberko, D. Ivaniuk, D. Shunkevich, O. Pupena Principles For Enhancing The Development And Use Of StandardsWithin Industry 4.0, Otkrytye semanticheskie tekhnologii proektirovaniya

intellektual'nykh system [Open semantic technologies for intelligent systems], 2020, pp. 167-174.

[11] A. Radford, K.Narasimhan, Improving Language Understanding by Generative Pre-Training, 2018

[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019 North American Chapter of the Association for Computational Linguistics.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020

[14] D. Xu, Wei C., W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, E. Chen Large Language Models for Generative Information Extraction: A Survey, 2020, arXiv preprint arXiv:2005.14165.

[15] A. Zagorskiy Integration Of Third-Party Functional Services On A Unified Semantic Basis, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 207-212.

[16] A. Cherkas, A. Kupo Ostis Technology Integration With Third-Party NLP Services, Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2023, pp. 207-212.

[17] "Introducing ChatGPT", Available at: https://openai.com/blog/chatgpt, (accessed 2023, January).

[18] P.Kohli, R.Gupta, S.Saxena, Robustness Challenges in Language Models: A Survey. arXiv preprint arXiv:2103.01021, 2021

[19] J. Dodge, S.Gururangan, D.Card, R.Schwartz, N. A.Smith, S. R.Bowman Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2020

[20] A. Holtzman, J. Buys, J. Du, M. Forbes, Y.Choi, The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751, 2020

[21] Y. Liu, M.Ott, N. Goyal, J. Du, M.Joshi, D.Chen, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692, 2019

[22] "T1-PLCnext-Demo", Available at: https://github.com/savushkin-r-d/T1-PLCnext-Demo, (accessed 2023, January).

NLP ПОДХОД К ПОСТРОЕНИЮ БАЗЫ ЗНАНИЙ ПРЕДПРИЯТИЯ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Таберко В. В., Иванюк Д. С.,
Смородин В. С., Прохоренко В. А.

Предложен подход к автоматизированному формированию базы знаний о предприятии на основании анализа существующей технической документации с применением больших языковых моделей. Данный подход позволяет обеспечить возможность интеграции предлагаемого решения с другими разработками, программными средствами предприятия для обеспечения построения интеллектуальных систем автоматизированного управления, рекомендательных систем и систем поддержки принятия решений, систем информационного обеспечения персонала предприятия.

Received 13.03.2024