Vol. 6, No. 1, 2024

Oleh Basystiuk[1], Nataliya Melnykova[2]

[1] Artificial Intelligence Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: oleh.a.basystiuk@lpnu.ua, ORCID 0000-0003-0064-6584

[2] Artificial Intelligence Systems Department, Lviv Polytechnic National University, Ukraine, Lviv, S. Bandery street 12, E-mail: nataliia.i.melnykova@lpnu.ua, ORCID 0000-0002-3257-3677

DEVELOPMENT OF THE MULTIMODAL HANDLING INTERFACE BASED ON GOOGLE API

Received: March 12, 2024 / Revised: March 28, 2024 / Accepted: April 01, 2024

https://doi.org/

Abstract. Today, Artificial Intelligence is a daily routine, becoming deeply entrenched in our lives. One of the most popular and rapidly advancing technologies is speech recognition, which forms an integral part of the broader concept of multimodal data handling. Multimodal data encompasses voice, audio, and text data, constituting a multifaceted approach to understanding and processing information. This paper presents the development of a multimodal handling interface leveraging Google API technologies. The interface aims to facilitate seamless integration and management of diverse data modalities, including text, audio, and video, within a unified platform. Through the utilization of Google API functionalities, such as natural language processing, speech recognition, and video analysis, the interface offers enhanced capabilities for processing, analysing, and interpreting multimodal data. The paper discusses the design and implementation of the interface, highlighting its features and functionalities. Furthermore, it explores potential applications and future directions for utilizing the interface in various domains, including healthcare, education, and multimedia content creation. Overall, the development of the multimodal handling interface based on Google API represents a significant step towards advancing multimodal data processing and enhancing user experience in interacting with diverse data sources.

Keywords: Speech-to-Text, speech recognition, sequence-to-sequence, machine learning, artificial intelligence.

Introduction

This study's main objective is to examine open API services that offer multimodal data conversion from audio to text. Also, we'll define and suggest a method for building unified user interfaces for multimodal audio-to-text conversion using recurrent neural networks. Artificial intelligence technology has spread and became easier to employ during the past ten years. Natural language processing combined with speech recognition is one of the most promising AI technologies. Future living will be heavily reliant on new speech recognition techniques and technologies because they drastically reduce communication time by using voice/audio instead of text. The majority of these features have been developed for the web and are accessible via an external API.

Google Cloud is one of the most well-known systems, so we'll base this study on it. List the main benefits and drawbacks of this strategy and library. Also, depending on the Google API, we will look into and discuss the time and calculation complexity of solutions.

Problem Statement

This topic is important because there are many tasks in which artificial intelligence can be used, especially in the field of communication. The scope of application for artificial intelligence is expanding. The creation of such systems is intriguing and encouraging for both science and industry, as study in this

area enables examination of the process of turning sound impulses into text and the identification of chances for process improvement and optimization. Companies are eager to invest in and incorporate artificial intelligence-based research into their products in order to achieve more advanced technologies. It gives them the chance to streamline their job with sound signals, regardless of the industry they are in—translation, journalism, communication, or management, etc.

The article's primary goal is to discuss the key phases of the construction of a machine translation strategy based on the Google API. The benefits and drawbacks of a number of methodologies, including rule-based, statistical, and neural network-based, are discussed. The most appropriate software technique and organizational structure for developing solutions for evaluating multimodal data.

Moreover, the two methods for numbering unstructured data were reviewed in terms of their software architecture and design. The translation of the produced recommended architecture approach into a full-size system and deployment to the market may be the next step in this research's development.

Main Material Presentation

The "artificial intelligence" industry [1] first emerged in the late 1950s, but its most active period of growth didn't start until the 2010s. Through the use of data structures like decision trees, graphs, Petri nets, and learning algorithms with datasets, scientists and developers can now create rules statically rather than beforehand.

The learning process itself is the process of developing a computer, for example, the same decision tree, but independently, after learning, the quality of recognition is evaluated using new data sets. Artificial intelligence is frequently used to simplify and automate specific activities, such as driving a car, facial recognition, natural language processing.

All these innovations have long entered people's lives and are sometimes employed without even noticing them. Natural language processing systems have replaced autopilot systems in this industry, which has already seen the use of facial recognition technology in social networks, security, and access control systems. journalism, translation, virtual assistants, and communication.

A wide range of industries have been affected by artificial intelligence, which helps to innovate, optimize, and in some cases completely replace human labor. Interest in these technologies, as a rule, grows every year, and the breadth of their usage widens, today artificial intelligence is employed in marketing, education, health care, games and many others [2].

Transfer systems work on a very simple principle: rules are applied to the incoming message that fit the structure of the outgoing message. In order to improve the internal representation of the information included in the message, the first stage of the task involves morphological, syntactic, and occasionally semantic analysis of the message. A translation is constructed from this representation utilizing multilingual dictionaries and grammatical rules. On the foundation of the primary representation that was derived from the original text, a more "abstract" internal representation may occasionally be constructed. This is done in order to highlight key conversion points and eliminate information that isn't necessary.

The internal representational levels are transformed in the opposite order as the translation text is constructed. When this strategy is used, translations of high quality are produced. The activity of any exchange transformation framework comprises of somewhere around five sections:

- morphological analysis;
- categorization of lexicons;
- lexical transfer;
- structure transfer;
- generation of morphology.
- The main functions of this system:
- receive the user's incoming message and process it;
- the received message is checked for audio;
- the received audio file is processed and converted into text;
- send the received result as a response to the user.

Let's take a closer look at each of the elements of this structure.

The package for processing and sending requests to the social network is contained in a single file (class), which contains functions for processing incoming messages, extracting audio files with messages, and generating responses to requests. A separate path has also been created to handle requests from browsers and present brief job information to interested users.

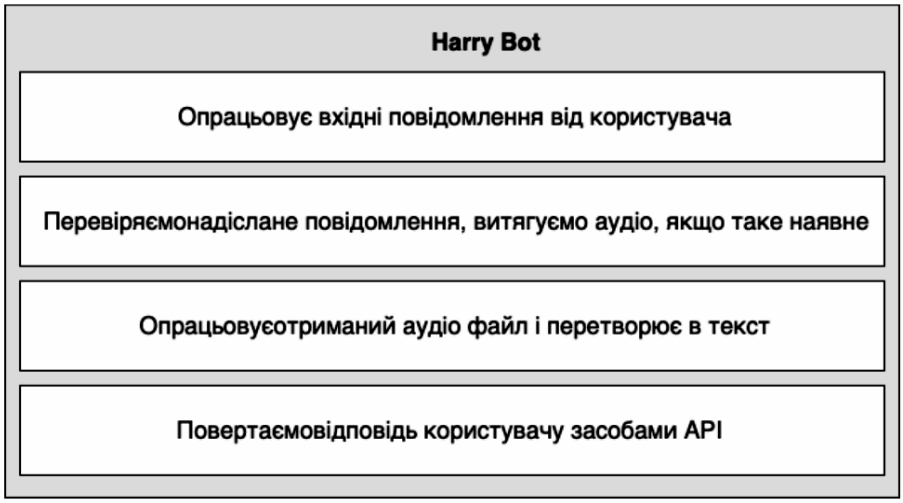

Fig. 1. Harry bot audio-to-text system features

Methods

In this section, we will describe the process of designing and developing the interface. We will explain the data collection and preprocessing techniques used to prepare the multimodal data for analysis. We will also describe the audio-to-text multimodal recognition module and the multimodal data interface module used to visualize and interact with the data.

The development of the multimodal handling interface is an important step towards facilitating the analysis and interpretation of complex multimodal data. It enables researchers and practitioners to combine data from multiple sources and modalities to gain deeper insights into complex phenomena. In the following sections, we will describe future system architecture approach, which consists of:

1. Multimodal Interface: The multimodal data interface module is a component of multimodal data analysis that provides a user-friendly interface for interacting with and visualizing multimodal data. It is designed to allow users to view, manipulate, and analyze data from multiple modalities simultaneously.

2. Preprocessing module: Preprocessing multimodal data module is an essential component of multimodal data analysis, which involves transforming raw multimodal data into a suitable format that can be used for further analysis. This module is designed to handle multiple modalities such as text, speech, images, videos, and gestures simultaneously.

3. Database connection: A database connection represent module where we establish a communication link between a database management system (DBMS) and a multimodal handling data interface. It is a vital aspect of any application that interacts with a database, allowing the application to access and manipulate data stored in the database.

4. Transformation interface: A request handling module is a component of a software application that handles incoming requests from clients and processes them to generate a response. It is typically part of a server-side application that receives requests from various clients and returns appropriate responses. In our case, that’s a bridge between database module and Google API, so in this module, we’ll prepare data to fit request requirements and validate them before sending.

5. Recognition module: The audio-to-text multimodal recognition module is a component of a multimodal system that is designed to convert audio signals to text. This module is typically used in

applications such as speech recognition, voice-to-text transcription, and audio annotation. In our case, we fully delegate this task to Google Cloud service, Audio-to-Text services via the Google API layer.

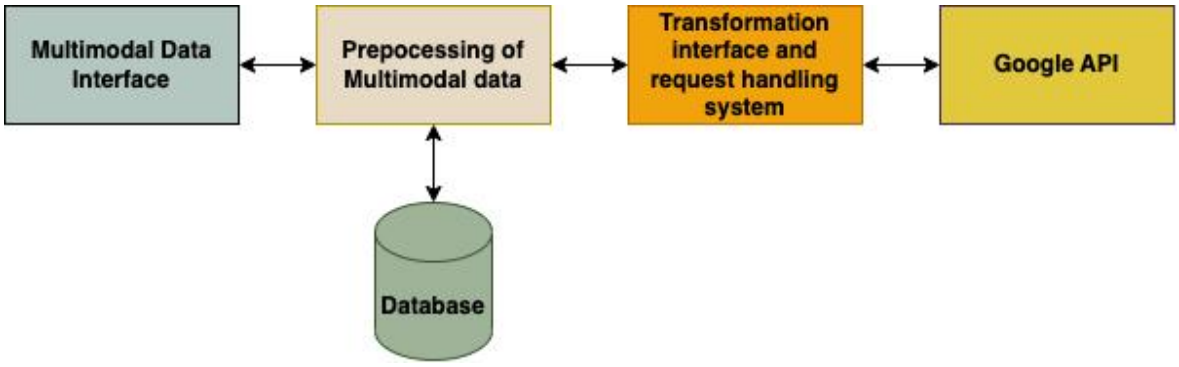

Fig. 2. Multimodal Handling Interface Based on Google API

The following are the method's primary advantages:

- The suggested methodology is constrained by the size of the training data set and the amount of processing resources that can be devoted to translation. Researchers of machine learning have established this method only a few years ago, but such systems are already operating better than the machine translation statistical systems, which were evolving through the last 20 years;
- The system does not depend on knowledge of any laws of the language. These guidelines are set forth by the algorithm itself and are updated frequently.

Results and Discussion

When you dive into natural language processing itself, we have sequence-to-sequence model of working with data, which is one of the most reliable and effective nowadays, as we already research in the previous articles and mentioned in Section 2. So current paper proposed an architecture solution for multimodal data processing, in field audio-to-text translation. In result, we'll present 2 main specifications of the proposed architecture:

- Deployment diagram;
- Flow diagram.

Deployment diagram describes how the process of software deployment on system components occurs. This diagram is most useful to people who are tasked with maintaining the designed system in the future, i.e. systems engineers, and it usually visualizes the performance, scalability, maintainability, and portability of the designed system. When the hardware components are displayed in relation to each other, it is easier to keep track of the entire hardware grid and ensure that all elements have been considered during the system deployment process, take a closer look at the component diagram in Figure 3;

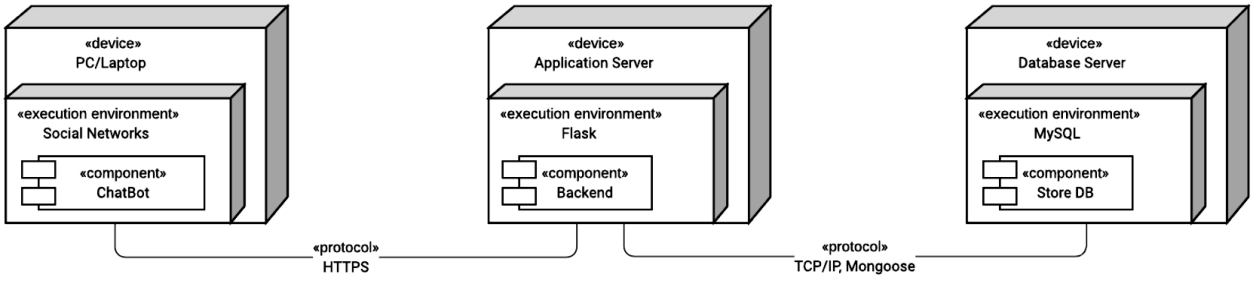

Fig. 3. Interface Deployment Diagram

The study involved several experiments to test the system's performance in different scenarios, including text-to-speech conversion, speech-to-text conversion, natural language processing, image recognition, and other machine learning-based services. The results of these experiments were analyzed and presented in this section.

The findings of this study provide valuable insights into the effectiveness of the multimodal handling interface based on Google API, highlighting its strengths and weaknesses. These results can be used to further refine and improve the system, as well as inform the development of similar multimodal interfaces in the future.

Overall, the results section of this article presents a comprehensive analysis of the performance of the multimodal handling interface based on Google API, providing valuable information for researchers and practitioners interested in developing similar systems.

So final approach architecture will be presented in this way:

1. Application Interface: that's a user-friendly data interface for the multimodal handling interface based on Google API, which provides a platform for handling and analyzing multimodal data.
2. Data handling interface: The Multimodal data interface module also facilitates data alignment and synchronization across different modalities.
3. Database connection: The database connection module includes several sub-modules that enable users to interact with the database, such as query builders, data mappers, and schema builders. These sub-modules simplify the process of creating and managing database tables and records, and enable users to perform complex queries and data analysis tasks. The database connection module also includes robust security features to protect the data stored in the database. Overall, the database connection is a crucial component of the multimodal handling interface based on Google API, enabling users to store, retrieve, and manage complex multimodal data efficiently and securely.
4. Google API interface: The Google API interface is designed to be easy to use, with detailed documentation and code samples available for developers to reference. It also provides a high level of security and reliability, with robust authentication and error handling mechanisms in place. Overall, the Google API interface is a powerful tool that allows developers to incorporate a range of Google services into their applications, providing users with a seamless and integrated experience.
5. Google Cloud function: this interface is a critical component of the multimodal handling interface, as it provides users with a comprehensive platform to handle and analyze complex multimodal data. It enables researchers and practitioners to gain deeper insights into complex phenomena by combining data from multiple sources and modalities.
6. Text-to-Speech API: component of Google Cloud API that converts text into natural-sounding speech. This API is designed to provide high-quality speech synthesis in a variety of languages and voices.
7. Google AutoML: component of Google Cloud, which provides deep neural networks (DNNs) to generate speech that will provide an option to post validate reply of the Google Cloud Text-to-Speech service API. The neural network model is trained on a large dataset of recorded human speech to learn the patterns of speech production and generate accurate and natural-sounding speech. That's module will help us to receive high quality result data, and to remove low quality results and ignore them in future, or reinforce this direction with adding correct data.

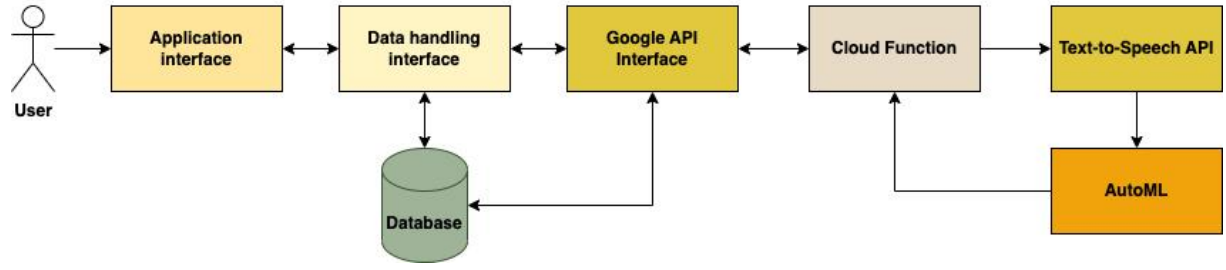

Fig. 4. Proposed architecture of audio-to-text system

Ultimately, it looks promising to use big data to build and train a machine learning model for automatic recognition. Reliability and quality of collected data is one of the main problems in big and multimodal data, in particular, that is why using third-party pre-trained libraries and arrays is a good way to solve natural language processing problems. As a result, a more finalized structure of the system was developed and proposed, which will provide an opportunity to obtain high-quality results of audio-to-text processing.

Also, when storing data, it is worth emphasizing the confidentiality and privacy of data. Because very often the information will come from private individuals, which may contain personal information, and we cannot neglect the quality of its storage.

The development of the Multimodal Handling Interface (MHI) based on Google API is a significant contribution to the field of human-computer interaction. This study aimed to design an efficient and intuitive interface that combines multiple modalities such as voice, touch, and gesture recognition to enhance user experience while interacting with a computing device.

In this study, we utilized the Google API to develop the MHI, which allowed us to incorporate various modalities, such as speech-to-text and natural language processing, into the interface. The design of the MHI is intuitive and easy to use, with minimal training required to interact with it.

However, there were some limitations to this study. One limitation was the sample size of the user study, which may not be representative of the entire population. Additionally, the study focused only on the use of the MHI on a smartphone, and it would be interesting to see how the interface performs on other computing devices such as tablets and laptops.

Overall, this study provides valuable insights into the development of multimodal interfaces and highlights the potential of using Google API to create efficient and intuitive interfaces that enhance user experience while interacting with computing devices.

Conclusions

In the process of conducting this research and preparing the article, we proposed a unified architecture approach, which could be utilized in development of multimodal audio-to-text recognition applications.

Today, one of the most well-liked areas of machine learning is natural language processing. This is mostly due, in my opinion, to the wide range of language processing related and multimodal handling applications. The paper presents scalable software solution for collecting and processing audio information to text. The proposed architecture created on Google Cloud API approach, with utilization of Audio-to-Text method and AutoML, for formalizing reply data. According to the results of research and was proposed a clear structure, which will be a basement to impalement and multimodal data conversion systems, especially contracted on audio to text conversations, in current research.

Our ongoing research aims to enhance the framework, starting with Section 6's recommendations to offer a stable infrastructure for the creation of third-party multimodal audio-to-text translation applications. We are aware of the necessity to add more interpretation qualities to the multimodal approach, which may be applied in a variety of disciplines, in order to supplement the detection of speech and other audio data. As a result, a new classification model that incorporates all of these traits will be created. Once this extension is complete, we might compare it to the current Google API approach and other TensorFlow or Keras-based options. The framework might also be a component of an autonomous system used in the sectors of medicine, journalism, entertainment, and communication. This system would support decision-making in these processes and assume more socially acceptable behavior toward people while reducing their workload.

Acknowledgements

In particular, it fit into the framework of research carried out in accordance with state funding at the Department of Artificial Intelligence Systems "Technologies for processing multimodal Ukrainian-language data to determine the level of stress" (State Register No. 0123U100231).

References

[1] Karpathy and L. Fei-Fei, “Deep visual-semantic alignmentsfor generating image descriptions,” in Proceedings of the IEEEComputer Society Conference on Computer Vision and PatternRecognition (CVPR), 2015, pp. 3128–3137 https://doi.org/10.1109/CVPR.2015.7298932

[2] Daxin Tan, Liqun Deng, Yu Ting Yeung, Xin Jiang, Xiao Chen,and Tan Lee, “Editspeech: A text based speech editing systemusing partial inference and bidirectional fusion,” arXiv preprintarXiv:2107.01554, 2021. https://doi.org/10.1109/ASRU51503.2021.9688051

[3] M. Oncescu, A. S. Koepke, J. F. Henriques, Z. Akata, andS. Albanie, “Audio Retrieval with Natural Language Queries,”in Proceedings of Conference of the International Speech Com-munication Association, 2021, pp. 2411–2415. https://doi.org/10.21437/Interspeech.2021-2227

[4] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and YoshuaBengio, Deep learning, vol. 1, MIT press Cambridge, 2016

[5] Ivan Izonin, et. al., "The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production", International Journal of Intelligent Systems and Applications (IJISA), Vol.10, No.9, pp.40-47, 2018. https://doi.org/10.5815/ijisa.2018.09.05

[6] Havryliuk, M., Dumyn, I., Vovk, O. (2023). Extraction of Structural Elements of the Text Using Pragmatic Features for the Nomenclature of Cases Verification. In: Hu, Z., Wang, Y., He, M. (eds) Advances in Intelligent Systems, Computer Science and Digital Economics IV. CSDEIS 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 158. Springer, Cham. https://doi.org/10.1007/978-3-031-24475-9_57

[7] Vitaly Yakovyna, Natalya Shakhovska, "Software failure time series prediction with RBF, GRNN, and LSTM neural networks", Procedia Computer Science 207(4):837-847, https://doi.org/10.1016/j.procs.2022.09.139

[8] Nataliya Shakhovska, et. al.: "The Developing of the System for Autimatic Audio to Text Conversion", IT&AS’2021: Symposium on Information Technologies and Applied Sciences, March 5–6, 2021, Bratislava, Slovak Republic.

[9] uxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, EricBattenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif ASaurous, “Style tokens: Unsupervised style modeling, controland transfer in end-to-end speech synthesis,” in InternationalConference on Machine Learning. PMLR, 2018, pp. 5180–5189.

[10] Nataliya Boyko, et. al.: "Usage of Machine-based Translation Methods for Analyzing Open Data in Legal Cases". In: Proc. of the CybHyg-2019, Kyiv, Ukraine, November 30, 2019, pp. 328–338. CEUR-WS.org.

[11] Berezsky O., Verbovyy S., Pitsun O. Hybrid Intelligent information techology for biomedical image processing. Proceedings of the IEEE International Conference «Computer Science and Information Technologies» CSIT’2018, Lviv. Ukraine, 11-14 September, 2018. P. 420-423. ïhttps://doi.org/10.1109/STC-CSIT.2018.8526711

[12] Zoryana Rybchak, et. al. "Analysis of methods and means of text mining". ECONTECHMOD, 6(2), 2017, pp. 73-78.

[13] P. Zdebskyi, V. Lytvyn,Y. Burov, and et. Intelligent system for semantically similar sentences identification and generation based on machine learning methods, CEUR Workshop Proceedings, 2020, pp. 317–346.

[14] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and MingLiu, “Neural speech synthesis with transformer network,” inProceedings of the AAAI Conference on Artificial Intelligence,2019, vol. 33, pp. 6706–6713. https://doi.org/10.1609/aaai.v33i01.33016706

[15] Oleh Basystiuk, Nataliia Melnykova "Multimodal Approaches for Natural Language Processing in Medical Data" Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine, Lyon, France, November 18 – 20, CEUR-WS.org, 2022. pp. 246-252

[16] N. Shakhovska, N. Boyko, P. Pukach. The Information Model of Cloud Data Warehouses International Conference on Computer Science and Information Technologies, CSIT 2018, September 11-14, Lviv, Ukraine, 2019, pp. 182-191. https://doi.org/10.1007/978-3-030-01069-0_13

[17] ifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng,“Phonetic posteriorgrams for many-to-one voice conversionwithout parallel data training,” in 2016 IEEE InternationalConference on Multimedia and Expo (ICME). IEEE, 2016, pp.1–6. https://doi.org/10.1109/ICME.2016.7552917

[18] S. Chowdhury and J. Sil, "FACERECOGNITION from NON-FRONTALIMAGES Using DEEP NEURALNETWORK," in 2017 Ninth InternationalConference on Advances in PatternRecognition (ICAPR), 2017, pp. 1-6. https://doi.org/10.1109/ICAPR.2017.8593160

[19] Z. Rybchak, O. Basystiuk, Analysis of computer vision and image analysis technics, ECONTECHMOD: an international quarterly journal on economics of technology and modelling processes, Lublin, Poland, 2017, pp. 79-84.

[20] I. Zheliznyak, Z. Rybchak, I. Zavuschak, Analysis of clustering algorithms, 2017. Advances in Intelligent Systems and Computing, 2017, pp. 305–314. https://doi.org/10.1007/978-3-319-45991-2_21

Development of the Multimodal Handling Interface Based on Google API

Олег Басистюк[1], Наталія Мельникова[2]

[1] Катедра систем штучного інтелекту, Національного університету "Львівська політехніка", Україна, Львів, вул. С. Бандери, 12, E-mail: oleh.a.basystiuk@lpnu.ua, ORCID 0000-0003-0064-6584
[2] Катедра систем штучного інтелекту, Національного університету "Львівська політехніка", Україна, Львів, вул. С. Бандери, 12, E-mail: nataliia.i.melnykova@lpnu.ua, ORCID 0000-0002-3257-3677

РОЗРОБКА ІНТЕРФЕЙСУ ОБРОБКИ МУЛЬТИМОДАЛЬНИХ ДАНИХ НА ОСНОВІ GOOGLE API

Отримано: березень 12, 2023 / Переглянуто: березень 28, 2023 / Прийнято: квітень 01, 2023

Анотація. Сьогодні штучний інтелект – це повсякденна рутина, яка глибоко увійшла в наше життя. Однією з найпопулярніших технологій, що швидко розвивається, є розпізнавання мовлення, яке є невід'ємною частиною ширшої концепції обробки мультимодальних даних. Мультимодальні дані охоплюють голос, аудіо та текстові дані, що є багатогранним підходом до розуміння та обробки інформації. У цій статті представлено розробку інтерфейсу для роботи з мультимодальними даними з використанням технологій Google API. Інтерфейс має на меті полегшити безперешкодну інтеграцію та управління різними форматами даних, включаючи текст, аудіо та відео, в рамках єдиної платформи. Завдяки використанню функцій Google API, таких як обробка природної мови, розпізнавання мови та аналіз відео, інтерфейс пропонує розширені можливості для обробки, аналізу та інтерпретації мультимодальних даних. У статті обговорюється дизайн і реалізація інтерфейсу, висвітлюються його особливості та функціональні можливості. Крім того, досліджуються потенційні застосування та майбутні напрямки використання інтерфейсу в різних сферах, включаючи охорону здоров'я, освіту та створення мультимедійного контенту. Загалом, розробка інтерфейсу для обробки мультимодальних даних на основі Google API є значним кроком на шляху до вдосконалення обробки мультимодальних даних та покращення користувацького досвіду взаємодії з різними джерелами даних.

Ключові слова: перетворення мови в текст, розпізнавання мови, sequence-to-sequence, машинне навчання, штучний інтелект.