DOI 10.31891/2307-5732-2024-335-3-2
УДК 004.62

БАСИСТЮК ОЛЕГ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0003-0064-6584
e-mail: oleh.a.basystiuk@gmail.com
МЕЛЬНИКОВА НАТАЛІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-2114-3436
e-mail: nataliia.i.melnykova@lpnu.ua
ДУМИН ІРИНА
Національний університет «Львівська політехніка»
https://orcid.org/0000-0001-5569-2647
e-mail: iryna.b.shvorob@lpnu.ua

РОЗРОБЛЕННЯ МУЛЬТИМОДАЛЬНОГО ІНТЕРФЕЙСУ НА ОСНОВІ GOOGLE API

В статті досліджується використання Google API для створення інноваційних мультимодальних інтерфейсів з метою покращення користувацького досвіду та продуктивності у різних сферах. Метою дослідження є розробка архітектурного підходу до обробки та аналізу мультимодальних даних. У дослідженні описано проектування та реалізацію інтерфейсу з використанням різних Google API для розпізнавання мови, обробки природної мови та розпізнавання жестів. Стаття також обговорює ключові етапи побудови стратегії машинного перекладу на основі Google API, визначає переваги та недоліки різних методологій та встановлює найбільш підходящі програмні техніки для розробки рішень для оцінювання мультимодальних даних. Два методи нумерації неструктурованих даних також розглядаються з точки зору їх програмної архітектури та дизайну. Запропонована система використовує сервіси Google Cloud Platform для надійного об'єднання даних з різних джерел та узагальнення їх у вихідні дані з високим коефіцієнтом розпізнавання успіху. Експерименти підтвердили доцільність використання мультимодального інтерфейсу обробки даних на основі Google API та описали його архітектурне рішення. Дослідження може бути використане для створення моделей перетворення мови в текст для конкретних медичних галузей, що покращить завдання перекладу мовлення та підвищить ефективність використання часу медичними працівниками. тканини.

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

BASYSTIUK OLEH, MELNYKOVA NATALIIA, DYMUN IRYNA
Lviv Polytechnic National University

DEVELOPMENT OF THE MULTIMODAL HANDLING INTERFACE BASED ON GOOGLE API

Context. In general, the article demonstrates the potential of Google APIs for developing innovative multimodal interfaces that can improve user experience and productivity in various fields.

Objective. The purpose of the work is to create an architectural approach to the processing and analysis of multimodal data.

Methods. This paper discusses the design and implementation of the interface, including the integration of various Google APIs for speech recognition, natural language processing, and gesture recognition. We also describe the technical details of the interface development, including the software and hardware components used, and provide examples of its use in real-world scenarios. In addition, a comparative analysis of existing approaches was carried out in order to select the most advanced and reliable ones for building reliable multimodal audio-to-text systems and conducting further research. The main purpose of the article is to discuss the key stages of building a machine translation strategy based on the Google API. The advantages and disadvantages of a number of methodologies are discussed, including rule-based, statistical, and neural network methodologies. The most appropriate software technique and organizational structure for developing solutions for evaluating multimodal data are identified. In addition, two methods for numbering unstructured data were considered in terms of their software architecture and design. The next step in the development of this research may be the implementation of the recommended architectural approach into a suitable system and its introduction to the market.

Results. The proposed system uses Google Cloud Platform services such as Speech-to-Text, Natural Language Processing, and AutoML to provide a reliable way to combine data from different sources and summarize it into relevant output with a success rate.

Conclusions. The experiments confirmed the feasibility of using a multimodal data processing interface based on Google API and described its architectural solution. This research was aimed at developing an effective and intuitive interface that combines several modalities, such as voice, touch, and gesture recognition, to improve the user experience when interacting with a computer device. This research can be used in the future to create a wide range of speech-to-text models for specific medical fields, which will improve the task of speech translation, reduce workload, and increase the efficiency of time use by medical professionals.

Keywords: speech-to-text conversion, speech recognition, Sequence-to-Sequence, machine learning, artificial intelligence.

Постановка проблеми

Основна мета статті – обговорити ключові етапи побудови стратегії машинного перекладу на основі Google API. Обговорюються переваги та недоліки низки методологій, зокрема, заснованих на правилах, статистичних даних і нейронних мережах. Визначено найбільш підходящу програмну техніку та організаційну структуру для розробки рішень для оцінювання мультимодальних даних.

Крім того, два методи нумерації неструктурованих даних були розглянуті з точки зору їх програмної архітектури та дизайну. Наступним кроком у розвитку цього дослідження може стати реалізація

рекомендованого архітектурного підходу у системі та виведення її на ринок.

Для досягнення цієї мети були визначені наступні основні завдання дослідження:

- Проаналізувати існуючі технології аналізу мультимодальних даних;
- Розробити метод обробки мультимодальних даних на основі Google Cloud;
- Огляд архітектури системи для аналізу мультимодальних даних.

Аналіз останніх джерел

Інновації вже давно увійшли в життя людей, і іноді їх використовують, навіть не помічаючи. Системи обробки природної мови прийшли на зміну системам автоматичного розпізнавання в цій галузі, яка вже бачила використання технології розпізнавання облич у соціальних мережах, системах безпеки та контролю доступу, журналістиці, перекладі, віртуальних асистентів та комунікації [1-3].

Широкий спектр галузей зазнав впливу штучного інтелекту, який допомагає впроваджувати інновації, оптимізувати, а в деяких випадках і повністю замінити людську працю. Інтерес до цих технологій, як правило, зростає з кожним роком, а широта їх використання розширюється, сьогодні штучний інтелект застосовується в маркетингу, освіті, охороні здоров'я, іграх та багатьох інших [4-6].

Системи перекладу працюють за дуже простим принципом: до вхідного повідомлення застосовуються правила, які відповідають структурі вихідного повідомлення. Для того, щоб покращити внутрішнє представлення інформації, що міститься в повідомленні, на першому етапі завдання виконується морфологічний, синтаксичний, а іноді й семантичний аналіз повідомлення. Переклад будується на основі цієї репрезентації з використанням багатомовних словників і граматичних правил. На основі первинної репрезентації, отриманої з оригінального тексту, іноді може бути побудована більш «абстрактна» внутрішня репрезентація. Це робиться для того, щоб виділити ключові точки перетворення та усунути непотрібну інформацію [7-9].

Рівні внутрішньої репрезентації трансформуються у зворотному порядку в процесі побудови тексту перекладу. При використанні цієї стратегії створюються переклади високої якості {10].

Діяльність будь-якої схеми трансформації при перекладі складається приблизно з п'яти етапів

- морфологічний аналіз;
- категоризація лексики;
- лексична передача;
- передача структури;
- генерація морфології.

Основні функції цієї системи:

- отримання вхідного повідомлення від користувача та його обробка;
- перевірка отриманого повідомлення на наявність звуку;
- обробка отриманого аудіофайлу та перетворення його в текст;
- надсилання отриманого результату у відповідь користувачеві.

Метою роботи є дослідження методів машинного навчання для обробки та аналізу мультимодальних даних на основі Google API.

Виклад основного матеріалу

Розробка інтерфейсу для обробки мультимодальних даних є важливим кроком на шляху до полегшення аналізу та інтерпретації складних мультимодальних даних. Він дозволяє дослідникам і практикам об'єднувати дані з різних джерел і модальностей, щоб отримати глибше розуміння складних явищ. У наступних розділах ми опишемо підхід до майбутньої системної архітектури, який складається і представлений на рис. 1:

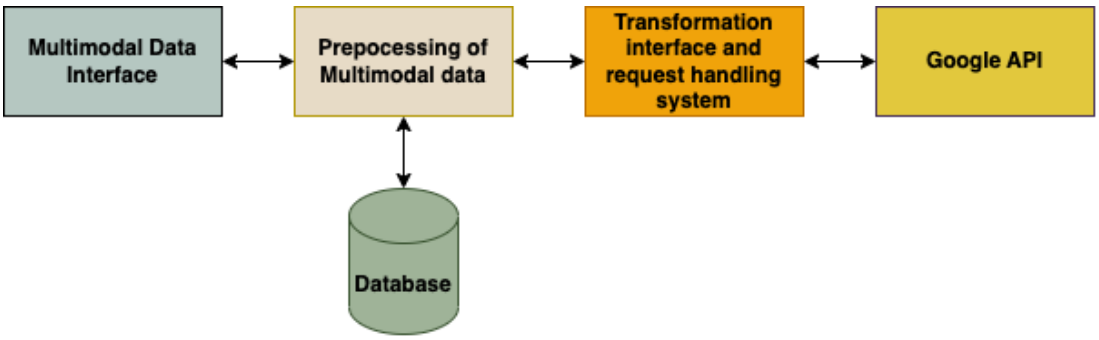

Рис. 1. Мультимодальний інтерфейс обробки на основі Google API

Запропонована система складається з таких модулів:

1. Мультимодальний інтерфейс: Модуль інтерфейсу обробки мультимодальних даних є компонентом аналізу мультимодальних даних, який забезпечує зручний інтерфейс для взаємодії та візуалізації мультимодальних даних. Він призначений для того, щоб дозволити користувачам переглядати, маніпулювати та аналізувати дані з різних видів транспорту одночасно.

2. Модуль попередньої обробки: Модуль попередньої обробки мультимодальних даних є важливим компонентом аналізу мультимодальних даних, який передбачає перетворення необроблених мультимодальних даних у відповідний формат, який може бути використаний для подальшого аналізу. Цей модуль призначений для одночасної обробки декількох модальностей, таких як текст, мова, зображення, відео та жести.
3. Підключення до бази даних: Модуль підключення до бази даних представляє собою модуль, за допомогою якого ми встановлюємо зв'язок між системою управління базами даних (СУБД) та інтерфейсом обробки мультимодальних даних. Це життєво важливий аспект будь-якої програми, яка взаємодіє з базою даних, дозволяючи програмі отримувати доступ до даних, що зберігаються в базі даних, і маніпулювати ними.
4. Інтерфейс перетворення: Модуль обробки запитів - це компонент програмного забезпечення, який обробляє вхідні запити від клієнтів та обробляє їх для формування відповіді. Зазвичай це частина серверної програми, яка отримує запити від різних клієнтів і повертає відповідні відповіді. У нашому випадку це місток між модулем бази даних і Google API, тому в цьому модулі ми підготуємо дані для повторних запитів і перевіримо їх перед відправкою.
5. Модуль розпізнавання: Модуль мультимодального розпізнавання аудіо-тексту - це компонент мультимодальної системи, який призначений для перетворення аудіосигналів у текст. Цей модуль зазвичай використовується в таких додатках, як розпізнавання мови, транскрипція голосу в текст і аудіокоментарі. У нашому випадку ми повністю делегуємо цю задачу хмарному сервісу Google, сервісу Audio-to-Text через рівень Google API.

Нижче наведено основні переваги цього методу:

1. Запропонована методологія обмежена розміром набору навчальних даних і кількістю обчислювальних ресурсів, які можуть бути виділені для перекладу. Дослідники машинного навчання створили цей метод лише кілька років тому, але такі системи вже працюють краще, ніж статистичні системи машинного перекладу, які розвивалися протягом останніх 20 років;
2. Система не залежить від знання будь-яких законів мови. Ці правила встановлюються самим алгоритмом і часто оновлюються.

Запропонована модель Google API для розробки інноваційних мультимодальних інтерфейсів, які можуть покращити користувацький досвід та продуктивність у різних сферах, наприклад, може підвищити ефективність використання робочого часу працівників.

Наприклад, запропонована система може використовувати сервіси хмарної платформи Google, такі як Speech-to-Text, Natural Language Processing та AutoML, щоб забезпечити надійний спосіб об'єднання даних з декількох джерел та узагальнення у відповідні вихідні дані з показником розпізнавання успіху. Ми розглянемо це посилення в наступному розділі.

Таким чином, остаточна архітектура підходу буде виглядати наступним чином і представлена на рис.2:

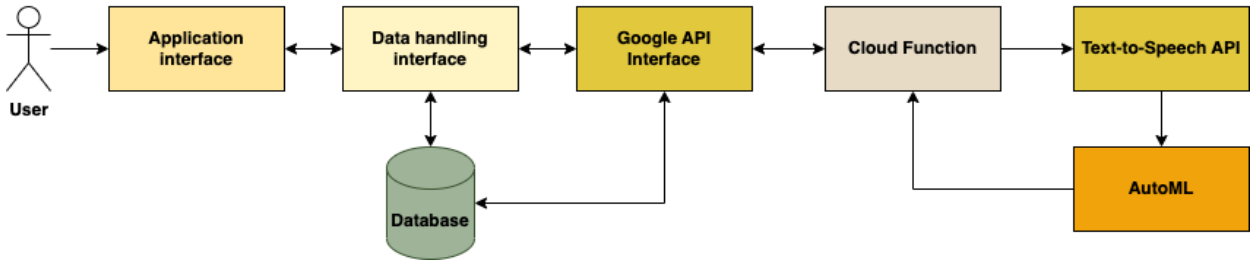

Рис. 2. Запропонована архітектура системи перетворення аудіо у текст

Зрештою, перспективним виглядає використання великих даних для побудови та навчання моделі машинного навчання для автоматичного розпізнавання. Надійність та якість зібраних даних є однією з головних проблем у великих та мультимодальних даних, зокрема, тому використання сторонніх попередньо навчених бібліотек та масивів є хорошим способом вирішення проблем обробки природної мови. В результаті було розроблено та запропоновано більш досконалу структуру системи, яка надасть можливість отримувати якісні результати обробки аудіо-тексту.

Також при зберіганні даних варто наголосити на конфіденційності та приватності даних. Адже дуже часто інформація буде надходити від приватних осіб, яка може містити персональні дані, і ми не можемо нехтувати якістю її зберігання.

Висновки

На сьогоднішній день одним з найбільш популярних напрямків машинного навчання є обробка природної мови. На мою думку, це здебільшого пов'язано з широким спектром програм, пов'язаних з обробкою мови та мультимодальною обробкою. У статті представлено масштабоване програмне рішення для збору та обробки аудіоінформації в текст. Запропонована архітектура створена на основі підходу Google

Cloud API, з використанням методу Audio-to-Text та AutoML для формалізації даних відповідей. За результатами дослідження було запропоновано чітку структуру, яка стане основою для впровадження мультимодальних систем перетворення даних, особливо орієнтованих на аудіо-текстові розмови, в подальших дослідженнях.

Наші поточні дослідження спрямовані на вдосконалення цієї структури, починаючи з результатів досліджень секцій та рекомендацій щодо їх обговорення, щоб запропонувати стабільну інфраструктуру для створення сторонніх додатків мультимодального перетворення аудіо-тексту в текст. Ми усвідомлюємо необхідність додати до мультимодального підходу, який може застосовуватися в різних дисциплінах, більше якостей інтерпретації, щоб доповнити розпізнавання мовлення та інших аудіоданих. В результаті буде створено нову класифікаційну модель, яка враховуватиме всі ці ознаки. Після завершення цього розширення ми зможемо порівняти його з поточним підходом Google API та іншими варіантами на основі TensorFlow або Keras. Фреймворк також може стати компонентом автономної системи, що використовується в галузях медицини, журналістики, розваг та комунікації. Така система підтримувала б прийняття рішень у цих процесах і передбачала б більш соціально прийнятну поведінку по відношенню до людей, зменшуючи при цьому їхнє робоче навантаження.

References

1. A. Karpathy, L. Fei-Fei, "Deep visual-semantic align-mentsfor generating image descriptions," in Proceedings of the IEEEComputer Society Conference on Computer Vision and PatternRecognition (CVPR), 2015, pp. 3128-3137.

2. D. Tan, L. Deng, Y. T. Yeung, X. Jiang, X. Chen, T. Lee, "Editspeech: A text-based speech editing systemusing partial inference and bidirectional fusion," arXiv pre-printarXiv:2107.01554, 2021.

3. M. Oncescu, A.S. Koepke, J.F. Henriques, Z. Akata, andS. Albanie, "Audio Retrieval with Natural Language Que-ries,"in Proceedings of Conference of the International Speech Com-munication Association, 2021, pp. 2411-2415.

4. I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, vol. 1, MIT press Cambridge, 2016

5. I. Izonin, A. Trostianchyn, Z. Duriagina, R. Tkachenko, T. Tepla, et. al., "The Combined Use of the Wiener Polyno-mial and SVM for Material Classification Task in Medi-cal Implants Production", International Journal of Intel-ligent Systems and Applications (IJISA), Vol.10, No.9, pp.40-47, 2018. DOI:10.5815/ijisa.2018.09.05

6. M. Havryliuk, I. Dumyn, O. Vovk, Extraction of Struc-tural 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, Com-puter 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. N. Shakhovska, V. Bilynska, O. Syvokon, O. Shamura-tov, 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.

8. I. Zheliznyak, Z. Rybchak, I. Zavuschak, Analysis of clus-tering algorithms, 2017. Advances in Intelligent Systems and Computing, 2017, pp. 305 314.

9. O. Basystiuk, N. Melnykova. MULTIMODAL SPEECH RECOGNITION BASED ON AUDIO AND TEXT DATA. Herald of Khmelnytskyi National University. 2022. Issue 5 (313). P. 22–25.

10. K. Shakhovska, I. Dumyn, N. Kryvinska, M. K. Kagita, "An Approach for a Next-Word Prediction for Ukrainian Language", Wireless Communications and Mobile Computing, vol. 2021, 2021. DOI: https://doi.org/10.1155/2021/5886119