Development of an ASR-Based Subtitle Generation System for Lecture Videos to Improve Searchability

Koichi Yoshizaki, Oita University, Japan

The Southeast Asian Conference on Education 2025
Official Conference Proceedings

Abstract
This study focused on improving the searchability of video-based online learning, specifically for Japanese lecture videos. A prototype semantic search API was developed to enhance search functionality using automatic speech recognition (ASR) and text embeddings. The system employs OpenAI Whisper to generate subtitles from uploaded videos. Text embeddings were generated using two models. The embeddings were stored in a vector database, enabling semantic search by calculating similarity between query embeddings and stored data. The system was evaluated on macOS using mlx-whisper, an optimized version of Whisper for Apple silicon. The preliminary evaluation demonstrated high ASR accuracy and efficient embedding generation, with the ruri-large model in particular providing more relevant search results for Japanese lecture videos.

Keywords: auto speech recognition, lecture video, text embedding, semantic search, web API

Introduction

Online learning frequently utilizes video content, which typically includes audio explanations accompanied by slides, computer screens, and other visual elements. However, the major limitation of video-based learning is the inability to easily search for specific information within the videos. This lack of searchability is a significant problem, especially when learning with lengthy videos or large collections of video files.

Recent advancements in machine learning have improved the accuracy of speech recognition software and made system implementation easier. Among these, Open AI Whisper, an open-model software, has demonstrated high recognition accuracy in many languages including Japanese, in addition to English (Radford et al., 2022). Furthermore, some automatic speech recognition (ASR) software like Whisper can preserve timestamps in transcription results and export subtitle data in formats such as SRT or VTT.

Document search methods include full-text search and semantic search, as well as hybrids of the two. The Japanese language poses unique challenges due to its three distinct scripts: Hiragana, Katakana, and Kanji. These scripts can represent the same word in different forms, leading to inconsistencies in written expressions. For this reason, the implementation of semantic search is important to realize highly accurate search for speech recognition results of videos using Japanese.

Objectives

The objectives of this study are to develop a Web API and vector database server with the following functions:

1. Automatically generate subtitle text data and the embeddings for uploaded videos.
2. Return relevant video IDs and time segments as search results for the query sentence sent to the API.
3. Evaluate server performance and search result accuracy for some lecture videos in Japanese.

Text embeddings are numerical representations of text that encode semantic meaning and contextual relationships between words, phrases, or documents. These embeddings transform textual data into high-dimensional vectors, enabling machines to process and analyze language effectively. In this study, a prototype of a semantic search API for videos, particularly those in Japanese, was designed, implemented, and evaluated. The system leverages text embeddings generated from the subtitle text of video content to implement semantic search functionality.

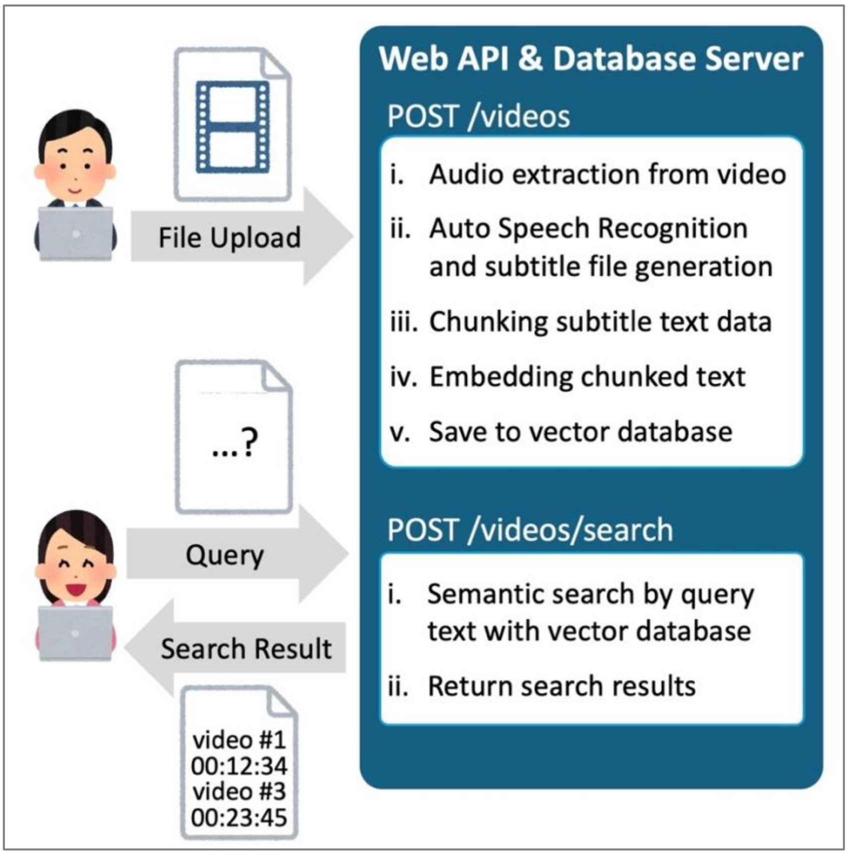
Proposed System

Using FastAPI, a Python framework, I developed a backend server consisting of a Web API and a vector database (Figure 1). This API system processes based on two operations via HTTP/HTTPS:

1. Uploading lecture video files by the instructor.
2. Inputting search terms or search queries by the learner.

When an instructor uploads a video file, the system extracts the audio track from the file and uses it as input for the ASR process. ASR and embeddings in this study are explained in the following subsections. It should be noted that this study did not develop a graphical user interface (GUI) to upload videos or perform semantic searches. Instead, all functionality was implemented and evaluated using a command-line interface (CUI).

Figure 1: The Concept of Proposed System


Web API & Database Server
File Upload
POST /videos
i. Audio extraction from video
ii. Auto Speech Recognition and subtitle file generation
iii. Chunking subtitle text data
iv. Embedding chunked text
v. Save to vector database
...?
Query
Search Result
POST /videos/search
i. Semantic search by query text with vector database
ii. Return search results
video #1
00:12:34
video #3
00:23:45

ASR: Auto Speech Recognition

This study utilizes OpenAI Whisper's "large-v3-turbo" model as speech recognition software (Radford et al., 2022). The "large-v3-turbo" model, one of multilingual models, was selected for its accuracy and speed balance. While Whisper provides the option to configure multiple parameters, this study specified only the model and the language (Japanese) settings. The main objective of this study is to perform a preliminary evaluation of the proposed system, so the "initial_prompt" parameter was not configured.

Although the output subtitle texts exhibited high accuracy, some misrecognitions were observed. These misrecognitions, along with character variations (Japanese has three writing systems), underscore the necessity of implementing semantic search functionality.

Embedding

Whisper's VTT output divides the recognition results into segments of appropriate length. In this study, some segments were merged into larger segments, ensuring that the total length did not exceed a specified maximum token size. The maximum token size parameter was carefully configured to maintain the semantic integrity of the subtitle text while allowing for the specification of playback positions within an appropriate time range during semantic search. Text embeddings were subsequently converted from the text in merged segments.

To generate text embeddings, two open-source embedding models were used: multilingual-e5-large (Wang et al., 2024) and ruri-large (Tsukagoshi & Sasano, 2024). The former model supports manylanguages including Japanese, whereas the latter model is especially specialized for Japanese. JMTEB, a benchmark for Japanese embedding models, reports results of STS (semantic textual similarity) tasks for various embedding models. Table 1 presents the JMTEB STS scores reported for the two embedding models utilized in this study, along with a comparison to a commercial model, OpenAI's text-embedding-3-large, which is widely used as a cloud-based service. In particular, the ruri-large model has demonstrated a high STS score for Japanese language.

Table 1: JMTEB STS Scores for Embedding Models (JMTEB, 2025)

Model	STS
cl-nagoya/ruri-large	83.13
OpenAI/text-embedding-3-large	82.52
intfloat/multilingual-e5-large	79.70

The text embeddings are stored in vector database, Chroma (Chroma, 2025). When saving text embeddings, the original text, playback time in the video, and video id were saved together as metadata. The embeddings stored in the vector database are utilized to process search queries entered by learners. When the search API is invoked with a term or phrase, an embedding is generated using the same model used to generate embeddings from the subtitle data. Semantic search is performed by calculating the similarity between this query embedding and the embeddings previously stored in the Chroma DB. The API returns multiple search results, ranked in descending order of similarity.

Evaluation

The proposed API system was evaluated for its functionality on a macOS device (Apple M4 Max, 16 cores, 64 GB RAM). The ASR system utilized in this evaluation was mlx-whisper, a version of the Whisper model optimized for Apple Silicon using the MLX framework. The API system was validated using several Japanese lecture videos, and the results confirmed its ability to achieve highly accurate automatic speech recognition and efficient embedding generation. End-to-end processing required approximately 3% of the video's duration.

While two embedding models produced similar similarity rankings for many search terms, the ruri-large model provided more appropriate results for certain queries.

Conclusion

The proposed system enables semantic search of video files using ASR-generated subtitles. Initial evaluations demonstrate highly accurate semantic search via a Japanese-optimized embedding model (Tsukagoshi & Sasano, 2024) as inferred from the JMTEB STS scores.

Since both embedding models have the same number of dimensions, there is no notable difference in terms of system load. Also, the system can run entirely in a local environment and can accommodate videos with non-public information.

Future work includes, first, the development of an API with a hybrid search function combining full-text search with semantic search. Additionally, the development and evaluation of a web-based user interface for searching and playing lecture videos using the API will be pursued.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21K12159.

References

Chroma. (2025). Chroma. https://www.trychroma.com

JMTEB. (2025). JMTEB Leaderboard. https://github.com/sbintuitions/JMTEB/blob/main/leaderboard.md

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356

Tsukagoshi, H., & Sasano, R. (2024). Ruri: Japanese General Text Embeddings. arXiv:2409.07737

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv:2402.05672

Contact email: kyoshi@oita-u.ac.jp