A Comparative Study on Enhancing the Accuracy of Chinese Speech-to-Text in Instructional Videos Using Large Language Models

Chih Chang Yang, National Taiwan Normal University, Taiwan
Tzren-Ru Chou, National Taiwan Normal University, Taiwan
Shu Wei Liu, National Taiwan University of Science and Technology, Taiwan

The Asian Conference on Education 2024
Official Conference Proceedings

Abstract

With the rapid development of speech recognition technology, Chinese speech-to-text (STT) systems play an important role in the production of subtitles and are often used in instructional videos. However, due to the complexity of the Chinese language and the large number of homophones, there is still significant room for improvement in the accuracy of existing STT systems. In this study, we proposed two optimization methods based on large language models (LLM), including language model-assisted editing and fine-tuned language model-assisted text editing, to improve the accuracy of Chinese STT, and verified them by producing subtitles for instructional videos in various domains and calculating the Levenshtein distance between two strings with dynamic programming. The results indicated that the fine-tuned language model-assisted text editing approach is significantly better than the language model-assisted editing approach in terms of text accuracy, and it can generate fine-tuning strategies for specific language characteristics to recognize language nuances more efficiently, thus significantly improving the accuracy of Chinese speech-to-text systems.

Keywords: Speech-to-Text (STT), Large Language Models (LLM), Instructional Videos, Fine-Tuned Language Models, Levenshtein Distance

iafor
The International Academic Forum
www.iafor.org

Introduction

In the modern era of increasingly prevalent digital education, Speech-to-Text (STT) technology has become a core tool for producing subtitles for instructional videos. However, due to the numerous homophones and complex grammatical structures in the Chinese language, existing STT systems still have significant room for improvement in accuracy (Chen et al., 2021; Zhang et al., 2018). Even advanced systems with multilingual recognition capabilities, such as OpenAI's Whisper and Google Cloud Speech-to-Text, perform suboptimally in handling Chinese, particularly in recognizing semantic differences (OpenAI, 2022; Wang, 2021).

The accuracy of subtitles in instructional videos directly affects learners' understanding of the content; incorrect word recognition may lead to misunderstandings of the learning material (Maraza-Quispe et al., 2022). Therefore, improving the accuracy of STT technology in Chinese environments has become an urgent issue to address. With the rise of Large Language Models (LLMs), which have the ability to understand textual content and make judgments, significant time can be saved on manual transcription (Brown et al., 2020; OpenAI, 2022). Although LLMs can handle tasks such as text organization, error detection, and correction, they still have limitations when dealing with homophones and specialized terminology (Maraza-Quispe et al., 2022).

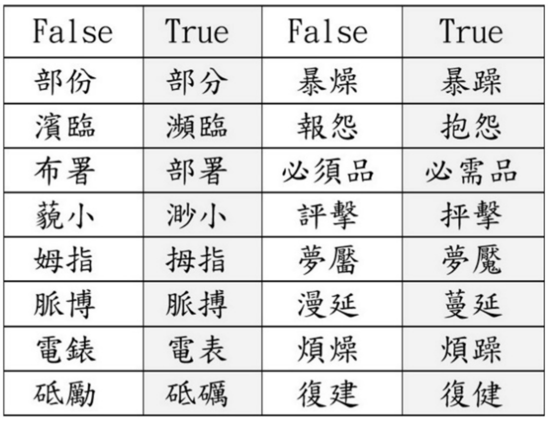
With the opening up of the fine-tuning functionality in LLMs, models can be optimized for cognitive abilities on specific data (OpenAI, 2023). After fine-tuning, the model can select and identify the correct words based on context; we believe it has the potential to improve the recognition of homophones and specialized terminology (Raffel et al., 2020).

Based on this, this study aims to explore and compare two Chinese STT optimization methods based on LLMs: language model-assisted editing (LMAE) and fine-tuned language model-assisted text editing. To compare the accuracy of the two processing methods, we use the Levenshtein distance calculated using dynamic programming algorithms to compute the minimum edit distance between strings, which measures the minimum number of edit operations required to transform one string into another—including insertion, deletion, and substitution of single characters—for[*] evaluation purposes (Che et al., 2017; Yujian & Bo, 2007).

Research Methodology

We first selected a sample of 60 Chinese instructional videos from higher education in fields such as humanities and social sciences, natural sciences, and engineering technology to simulate real-world application scenarios. Subsequently, we used existing STT systems (e.g., OpenAI Whisper) to generate the initial subtitles. Then, we employed a Large Language Model (LLM) to perform text organization, error detection, and correction on the initial subtitles; this process constitutes the language model-assisted editing.

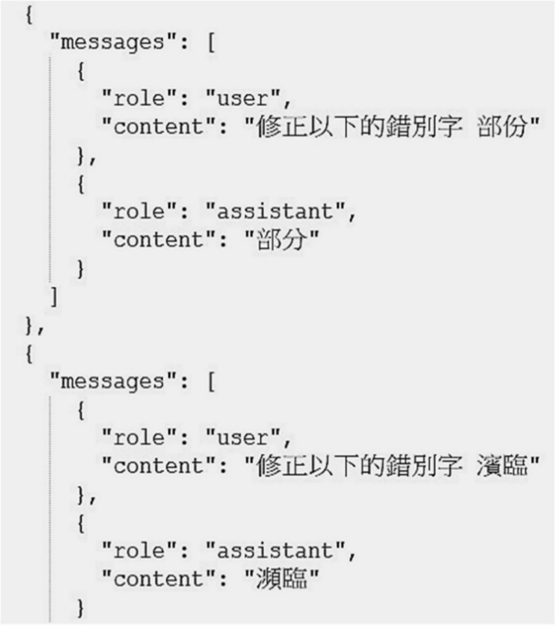
Next, for the fine-tuned language model-assisted text editing, we collected commonly used Chinese homophones and compiled various homophone tables or documents. We utilized the ChatGPT-4 multimodal model to identify and organize this information into the dialogue format required for fine-tuning, in JSONL format.



False	True	False	True
部份	部分	暴燥	暴躁
濱臨	瀕臨	報怨	抱怨
布署	部署	必須品	必需品
藐小	渺小	評擊	抨擊
姆指	拇指	夢壓	夢魘
脈博	脈搏	漫延	蔓延
電錶	電表	煩燥	煩躁
砥勵	砥礪	復建	復健

Homophones refer to Chinese characters that have identical phonetic forms but completely different character forms and meanings. For example, "部屬" (subordinates) and "部署" (deployment) are pronounced the same, but the former refers to personnel arrangements, while the latter refers to the arrangement of matters.

Figure1: Corrections of Common Homophones

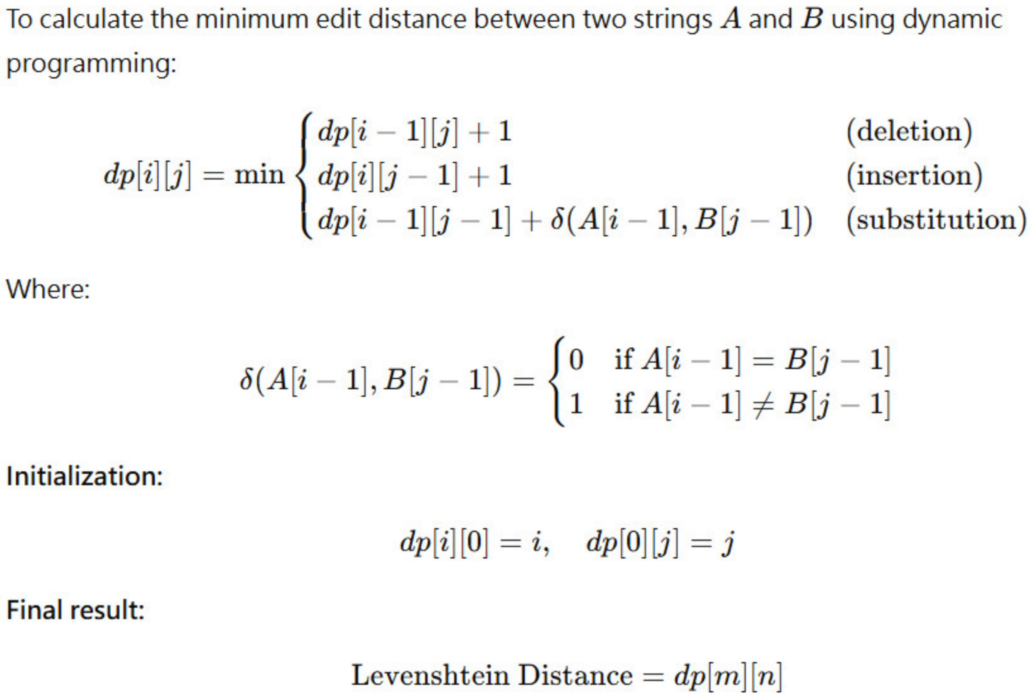


```
{
  "messages": [
    {
      "role": "user",
      "content": "修正以下的錯別字 部份"
    },
    {
      "role": "assistant",
      "content": "部分"
    }
  ]
},
{
  "messages": [
    {
      "role": "user",
      "content": "修正以下的錯別字 濱臨"
    },
    {
      "role": "assistant",
      "content": "瀕臨"
    }
```

Figure 2: JSONL File in Dialogue Format Required for Fine-Tuning

Using the organized JSONL file, we fine-tuned the ChatGPT-4o-mini model. Subsequently, we used the fine-tuned language model to assist in text editing, performing text organization, error detection, and correction on the initial subtitles.

To evaluate the results, we calculated the Levenshtein distance using dynamic programming algorithms to compute the minimum edit distance between strings. We first compared the shortest distances between each method and the expert-approved standard examples.



To calculate the minimum edit distance between two strings A and B using dynamic programming:

$$dp[i][j] = \min \begin{cases} dp[i-1][j] + 1 & \text{(deletion)} \\ dp[i][j-1] + 1 & \text{(insertion)} \\ dp[i-1][j-1] + \delta(A[i-1], B[j-1]) & \text{(substitution)} \end{cases}$$

Where:

$$\delta(A[i-1], B[j-1]) = \begin{cases} 0 & \text{if } A[i-1] = B[j-1] \\ 1 & \text{if } A[i-1] \neq B[j-1] \end{cases}$$

Initialization:

$$dp[i][0] = i, \quad dp[0][j] = j$$

Final result:

$$\text{Levenshtein Distance} = dp[m][n]$$

Figure 3: Levenshtein Distance Calculation Formula

A Levenshtein distance value closer to zero indicates fewer changes, signifying a closer match to the expert-approved standard examples. To assess the statistical significance of these results, we chose to use an independent samples t-test for statistical analysis.

Results

We conducted an accuracy evaluation of two methods: fine-tuned language model-assisted text editing and language model-assisted editing. An independent samples t-test was used to compare the accuracy differences between the two datasets. The results showed a significant difference between the two groups (t=2.65544, p=.004507). This indicates that the fine-tuned language model performs significantly better in Chinese speech-to-text tasks than the general language model-assisted editing method.

In calculating the relevant statistical data, the mean edit distance of the fine-tuned language model-assisted text editing group was lower (M=674.2), while that of the language model-assisted editing group was higher (M=858.53). This suggests that the fine-tuned language model more effectively handled challenges such as homophones and specialized terminology, significantly reducing instances of erroneous transcription.

Table 1: Descriptive Statistics and t-Test Results for Two Treatment Groups

Group	N	M	SD	t	*p*
language model-assisted editing	60	858.53	171.04	2.66	.005*
fine-tuned language model-assisted text editing	60	674.2	118.08		

$^{*}p < .05$

Conclusion

This study successfully demonstrated the effectiveness of fine-tuned language models in improving the accuracy of subtitles in Chinese instructional videos. Through our collected dataset of homophones and fine-tuning, the fine-tuned model exhibited higher language understanding and text generation capabilities, effectively overcoming the shortcomings of existing STT systems in accurately recognizing homophones.

The research results indicate that the fine-tuned language model can significantly reduce the error rate in the subtitle production process, further enhancing the quality and efficiency of subtitles in instructional videos. This study provides an alternative solution for improving the accuracy of Chinese speech-to-text technology and lays a foundation for subsequent applications in a wider range of educational contexts.

Future work will focus on expanding the fine-tuning dataset on a larger scale, targeting specialized terminology in various professional fields to further enhance the model's adaptability to diverse language scenarios. We will also explore how to apply this technology to other languages and domains, promoting the comprehensive development of speech recognition technology.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Che, X., Luo, S., Yang, H., & Meinel, C. (2017, July). Automatic lecture subtitle generation and how it helps. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)* (pp. 34-38). IEEE.

Chen, Y. C., Cheng, C. Y., Chen, C. A., Sung, M. C., & Yeh, Y. R. (2021). Integrated semantic and phonetic post-correction for chinese speech recognition. *arXiv preprint arXiv:2111.08400.*

Maraza-Quispe, B., Alejandro-Oviedo, O. M., Fernandez-Gambarini, W. C., Cuadros-Paz, L. E., Choquehuanca-Quispe, W., & Rodriguez-Zayra, E. (2022). Analysis of the cognitive load produced by the use of subtitles in multimedia educational material and its relationship with learning. *International Journal of Information and Education Technology*, *12*(8), 732-740.

OpenAI. (2022, September 21). *Whisper: Robust speech recognition via large-scale weak supervision*. OpenAI. https://openai.com/blog/whisper/

OpenAI. (2022, November 30). *ChatGPT: Optimizing language models for dialogue*. OpenAI. https://openai.com/blog/chatgpt/

OpenAI. (2023). *Fine-tuning language models*. OpenAI. https://platform.openai.com/docs/guides/fine-tuning

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Wang, H. H. (2021). Speech recorder and translator using Google cloud speech-to-text and translation. *Journal of IT in Asia*, *9*(1), 11-28.

Wilken, P., Georgakopoulou, P., & Matusov, E. (2022). SubER: A Metric for Automatic Evaluation of Subtitle Quality. *arXiv preprint arXiv:2205.05805*.

Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 1091-1095.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E. D., Jin, W., & Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *9*(5), 1-28.