Combining Large Language Model with Speech Recognition System in Low-resource Settings

Sheng Li[1] Zhengdong Yang[1,2] Wangjin Zhou[2] Chenhui Chu[2]
Chen Chen[3] Eng Siong Chng[3] Hisashi Kawai[1]
[1]National Institute of Information and Communications Technology (NICT)
[2]Kyoto University
[3]Nanyang Technological University
{sheng.li,hisashi.kawai}@nict.go.jp
zd-yang@nlp.ist.i.kyoto-u.ac.jp,zhou@sap.ist.i.kyoto-u.ac.jp,chu@i.kyoto-u.ac.jp
chen1436@e.ntu.edu.sg, ASESChng@ntu.edu.sg

概要

This paper investigates integrating automatic speech recognition (ASR) with large language models (LLMs). The overarching goal of the paper is to validate the effectiveness of integrating LLMs with ASR systems, with a specific focus on low-resource settings. In our experiment, The LLM is utilized for second-pass rescoring to correct errors in ASR outputs. We fine-tune a Japanese-specific version of the LLaMA model, named japanese-Llama-2-7b, feeding it with the Whisper-Large-v3 ASR model's n-best output. The experiment shows that the proposed method effectively enhances the ASR result, even in low-resource environments.

1 Introduction

Nowadays, methods of combining automatic speech recognition (ASR) applications with pre-trained language models (LMs) are booming.

Zhang et al. [1] proposed a spelling corrector based on the transformer [2] to reduce the substitution error in Mandarin speech recognition. [3] improved the BERT [4] effectiveness in detecting spelling errors with the soft-masking technique as the bridge between the error detector and corrector. Futami et al. [5] generated soft labels for ASR training with the BERT distilling knowledge. There are also some works [6, 7] studying to improve ASR rescoring by BERT. Additionally, BERT has also been successfully applied in multi-modal studies in vision-language pretraining [8, 9, 10, 11] or voice-language pretraining [12, 13, 14]. More recently, [15] proposes combining large language models (LLMs) into a speech recognition system.

This paper follows the rescoring method described in [15] and tests this method for Japanese ASR tasks and low-resource settings. It also extends our previous work combining BERT and GPT2 [16] with Wav2Vec2.0 ASR system [17].

2 Related Work

2.1 AMs for ASR task

Conventional hybrid Gaussian mixture and hidden-Markov (GMM-HMM) [18] and deep neural network and hidden-Markov (DNN-HMM) [19] based automatic speech recognition (ASR) systems require independently optimized components: acoustic model, lexicon and language model. The end-to-end (E2E) model integrates these components into a single neural network. It simplifies ASR system construction, solves the sequence labeling problem between variable-length speech frame inputs and label outputs (phone, character, syllable, word, etc.), and has achieved promising results on ASR tasks. Various types of E2E models have been studied in recent years: connectionist temporal classification (CTC) [20, 21], attention-based encoder-decoder (Attention) E2E models [22, 23], E2E lattice-free maximum mutual information (LFMMI) [24], and E2E models jointly trained with CTC and attention-based objectives (CTC/Attention) [25, 26].

The transformer has been applied to E2E speech recognition systems [27, 28, 29, 30] and has achieved promising

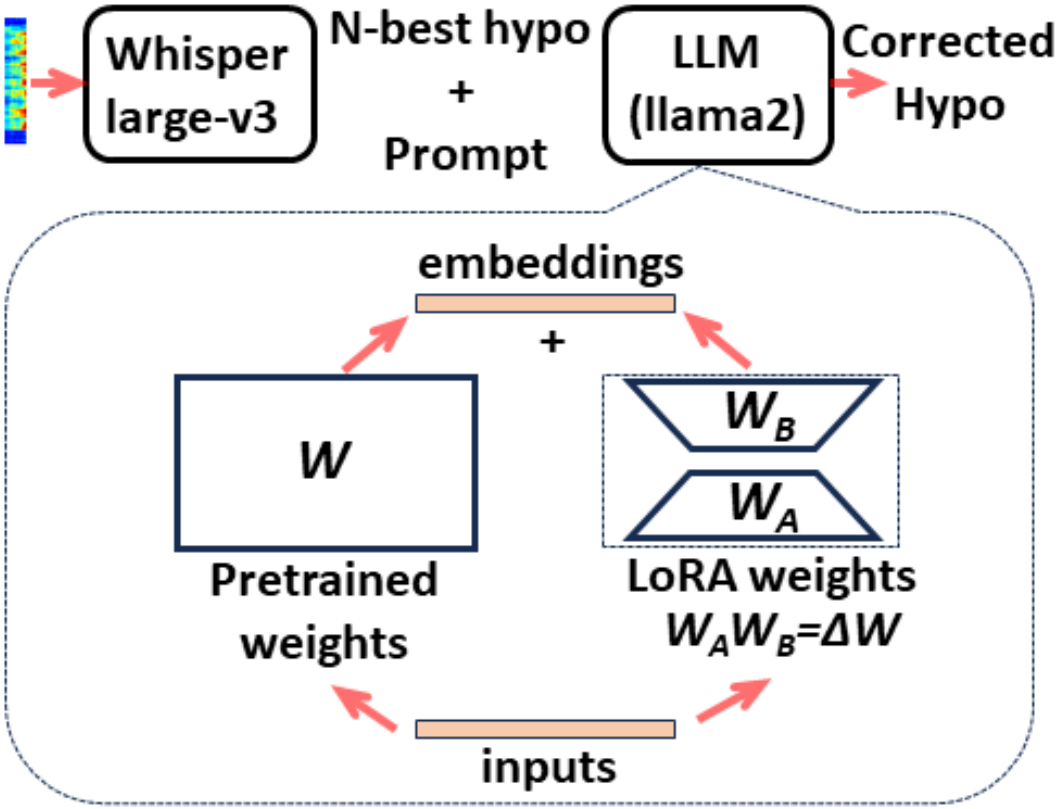


図 1: The flowchart of the method.

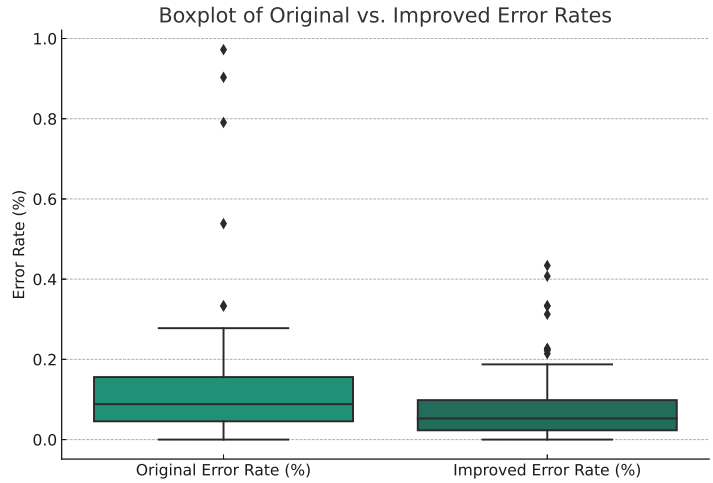
results. Then, the self-supervised learning (SSL) model, e.g., the Wav2Vec2.0 [31], became popular. The training is based on self-supervised learning with unlabeled speech. Then, the model is fine-tuned on labeled data with the CTC objective for the ASR task. Whisper [32], OpenAI's transformer-based open-source ASR system, excels in accuracy and contextual understanding. Trained on a vast 680,000-hour dataset of diverse languages and accents, it effectively transcribes in noisy environments. It is ideal for real-time applications like live captioning and voice-to-text conversion. In this paper, we use the Whisper model for our experiment.

2.2 LMs for ASR task

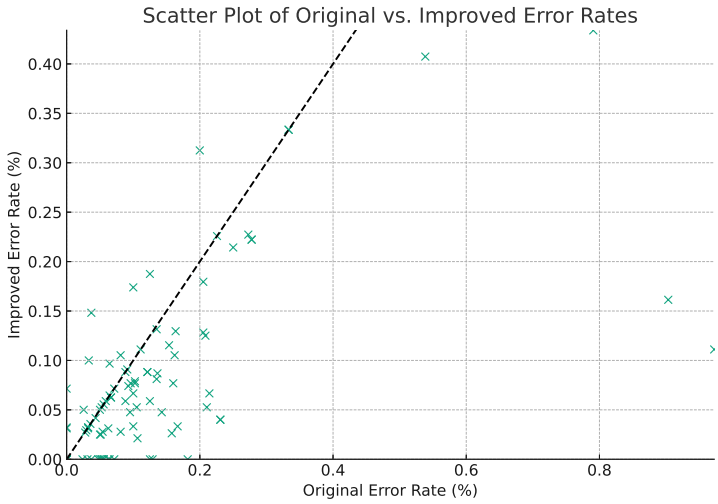
Using an LM in ASR brings a leap to speech recognition performance. Generally speaking, ASR combing with LM has two types of strategies: first-pass decoding and second-pass rescoring.

In ASR, the task was formulated as a noisy channel model using the Bayes rule $P(W|X) = P(X|W)P(W)$, where X is the speech signal and W is the corresponding text. The two distributions of $P(X|W)$ and $P(W)$ were named acoustic and language models, respectively. The LM was trained separately on the source text and only used for decoding [33]. WFST-based decoding compiles n-gram LMs into the decoding graph for efficient first-pass decoding [34]. Incorporating larger n-gram LMs made the decoding graph explode, and researchers changed the compiling improved algorithm [35] or used second-pass rescoring in both offline and on-the-fly settings [36, 37] instead. The Bayesian formulation still made sense in the hybrid DNN-HMM model era. The scores could be interpreted as pseudo-likelihoods by subtracting an appropriate prior, so the same decoding/rescoring framework carried over.

For End-to-End ASR, the models directly estimate $P(W|X)$. We still can combine using LMs in the first-pass decoding (e.g., shallow fusion, cold fusion, etc. [38, 39]).



(a) The boxplot analysis of error rates.



(b) The scatterplot analysis of error rates.

Similar to the second-pass rescoring, hypotheses can be obtained using beam search on an ASR model and re-rank with an externally trained LM. A two-pass E2E ASR model was proposed with an encoder shared between a streaming RNN-T model and a full-context LAS decoder [40]. There are also some works [1, 3, 5, 6, 7] studying to improve ASR rescoring by transformer and BERT. More recently, [15] proposes combining LLMs into a speech recognition system.

3 Method

This paper uses LLM for error correction, which is second-pass rescoring in the output transcriptions generated by the ASR system (N-best decoding hypotheses), as shown in Figure 1.

We introduce LoRA [41] to avoid tuning the whole set of parameters of a pre-trained model by inserting a neural module with a small number of extra trainable parameters to approximate the full parameter updates, allowing for efficient learning of the N-best to transcription mapping without affecting the pre-trained parameters of the LLM. Our method introduces trainable low-rank decomposition matrices into LLMs' existing layers, enabling the model to adapt to new data while keeping the original LLMs fixed to retain the previous knowledge. Specifically, LoRA performs a reparameterization of each model layer

expressed as a matrix multiplication by injecting low-rank decomposition matrices 1. As a result, the representations generated by the LLM are not distorted due to task-specific tuning. At the same time, the adapter module acquires the capability to predict the true transcription from the N-best hypotheses. Benefiting from efficient training, we can employ a large-scale language model in the method, which is expected to understand the task description and capture correlation in the N-best list.

4 Experiments

4.1 Experimental Settings

The experimental settings for fine-tuning a Japanese language model in a low-resource environment are as follows.

1. LLM used: The experiment employs the japanese-Llama-2-7b model[1)]. This model is presumably a variant of the LLaMA (Large Language Model by Meta AI) adapted for Japanese language processing.
2. ASR model: Whisper-Large-v3 generates 10-best outputs (CER%=12.91 for 1-best).
3. The training is performed on an NVIDIA Tesla V100 GPU using 8-bit training. The hyperparameters for finetuning are 15 epochs, learning rate 1e-4, batch size 64, and LoRA rank 4.
4. Dataset: To simulate low-resource settings, we use the Japanese dataset from SPREDS-U1 [2)]. The fine-tuning process involves 900 Japanese sentences for the low-resource setting. A separate 100 Japanese sentences for evaluation.
5. Evaluations: We use NIST-SCTK to evaluate the Character Error Rate (CER%).

4.2 Experimental Results

Table 1 shows the results of the experiments.

表 1: Evaluation of LLM-based Correction Model (CER%), the ASR output CER% is 12.91

ASR	Epoch 7	Epoch 9	Epoch 11	Epoch 13
12.91	26.20	8.24	**7.77**	16.51

We conduct the paired sample t-test on the results from the best model (Epoch 11). The paired sample t-test suggests that the improvement in CER% is statistically significant.

In Figure 2a, the left boxplot represents the Original CER%, and the right means the Improved CER%. This plot illustrates the results' spread and central tendency, where you can see the generally lower spread and median in the Improved CER%.

Each point in Figure 2b represents a pair of original and improved CER%. The dashed line represents the line of equality (where the original and improved rates would be equal). Points below this line indicate instances where the improved CER% is lower (better) than the original, and points above the line indicate the opposite. The concentration of points below the line further supports the conclusion that the improved CER% are generally lower than the original rates.

4.3 Further Discussions

The method we use in this paper can be further improved in the following aspects.

1. We notice that the Whisper-large model has an excessively high accuracy on the LibriSpeech-clean-test dataset. If the clean-test Word Error Rate (WER) is lower than 2%, using a LLM for correction is impossible. However, the proposed method works again when we switch to a relatively weak model, e.g., Whisper-tiny. Librispeech-other-test WER% reduced from 27.5% to 20.5%, and Librispeech-clean-test WER% reduced from 27.3% to 21.4%.
2. The original llama model cannot be directly used. Before we fine-tune the LLM model with N-best and ground truth data pairs, the LLM must be pretrained with large-scale language-specific textual data. However, there are no existing high-quality LLMs for low-resourced languages. That is why it is a pity we did not work on real low-resource languages.

5 Conclusion

This paper proves combining LLM with a speech recognition system effectively improves speech recognition performance, even in low-resource settings. In the future, we will conduct more experiments to demonstrate the method's effectiveness in a broader range of settings.

1) huggingface.co/elyza/ELYZA-japanese-Llama-2-7b
2) ast-astrec.nict.go.jp/en/release/SPREDS-U1

謝辞

This work was supported by JSPS KAKENHI Grant Numbers JP23K11227, JP23H03454, and NICT international funding.

参考文献

[1] Shiliang Zhang, Ming Lei, and Zhijie Yan. Investigation of transformer based spelling correction model for ctc-based end-to-end mandarin speech recognition. In **Proc. Interspeech**, pp. 2180–2184, 2019.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In **CoRR abs/1706.03762**, 2017.

[3] Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. Spelling error correction with soft-masked bert. **arXiv preprint arXiv:2005.07421**, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1 (Long and Short Papers)**, pp. 4171–4186, June 2019.

[5] Hayato Futami, Hirofumi Inaguma, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. Distilling the knowledge of BERT for sequence-to-sequence ASR. **CoRR**, Vol. abs/2008.03822, , 2020.

[6] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. **arXiv preprint arXiv:1910.14659**, 2019.

[7] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. Effective sentence scoring method using bert for speech recognition. In **Proc. ACML**, pp. 1081–1093, 2019.

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. **arXiv preprint arXiv:1908.02265**, 2019.

[9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. **arXiv preprint arXiv:1908.03557**, 2019.

[10] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In **Proc. AAAI**, Vol. 34, pp. 13041–13049, 2020.

[11] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In **Proc. ECCV**, pp. 121–137, 2020.

[12] Alexei Baevski and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for asr. In **2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, pp. 7694–7698, 2020.

[13] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In **Proc. IEEE-ICASSP**, pp. 6533–6537, 2021.

[14] Chengyi Wang, Yu Wu, Shujie Liu, Ming Zhou, and Zhenglu Yang. Curriculum pre-training for end-to-end speech translation. **arXiv preprint arXiv:2004.10093**, 2020.

[15] Cheng Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng Siong Chng. Hyporadise: An open baseline for generative speech recognition with large language models. **ArXiv**, Vol. abs/2309.15701, , 2023.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[17] Sheng Li and Jiyi Li. Correction while recognition: Combining pretrained language model for taiwan-accented speech recognition. In Lazaros Iliadis, Antonios Papaleonidas, Plamen Angelov, and Chrisina Jayne, editors, **Artificial Neural Networks and Machine Learning – ICANN 2023**, pp. 389–400, Cham, 2023. Springer Nature Switzerland.

[18] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, Vol. 77, No. 2, pp. 257–286, 1988.

[19] G. Dahl, D. Yu, L. Deng, and A. Acero. Context dependent pre-trained deep neural networks for large vocabulary speech recognition. **IEEE Trans. ASLP**, Vol. 20, No. 1, pp. 30–42, 2012.

[20] A. Graves and N. Jaitly. Towards End-to-End speech recognition with recurrent neural networks. In **Proc. ICML**, 2014.

[21] Y. Miao, M. Gowayyed, and F. Metze. EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding. In **Proc. IEEE-ASRU**, pp. 167–174, 2015.

[22] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In **Proc. NIPS**, 2015.

[23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In **Proc. IEEE-ICASSP**, 2016.

[24] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur. End-to-end speech recognition using lattice-free mmi. In **Proc. INTERSPEECH**, 2018.

[25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. **IEEE Journal of Selected Topics in Signal Processing**, Vol. 11, No. 8, pp. 1240–1253, 2017.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. Espnet: End-to-end speech processing toolkit. In **Proc. INTERSPEECH**, 2018.

[27] L. Dong, S. Xu, and B. Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In **Proc. IEEE-ICASSP**, 2018.

[28] S. Zhou, S. Xu, and B. Xu. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. In **CoRR abs/1806.05059**, 2018.

[29] S. Zhou, L. Dong, S. Xu, and B. Xu. A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese. In **CoRR abs/1805.06239**, 2018.

[30] S.Zhou, L.Dong, S.Xu, and B.Xu. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. In **Proc. INTERSPEECH**, 2018.

[31] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In **Advances in Neural Information Processing Systems (NeurIPS)**, Vol. 33, pp. 12449–12460, 2020.

[32] OpenAI. Whisper: Robust speech recognition via large-scale weak supervision. https://openai.com/blog/whisper/, 2023.

[33] Frederick Jelinek. Continuous speech recognition by statistical methods. **Proceedings of the IEEE**, Vol. 64, pp. 532–556, 1976.

[34] Mehryar Mohri, et al. Speech recognition with weighted finite-state transducers. 2008.

[35] Paul R. Dixon, Chiori Hori, and Hideki Kashioka. A specialized WFST approach for class models and dynamic vocabulary. In **Proc. Interspeech 2012**, pp. 1075–1078, 2012.

[36] Andrej Ljolje, et al. Efficient general lattice generation and rescoring. In **EUROSPEECH**, 1999.

[37] Hasim Sak, et al. On-the-fly lattice rescoring for real-time automatic speech recognition. In **Interspeech**, 2010.

[38] Jan Chorowski and Navdeep Jaitly. Towards better decoding and language model integration in sequence to sequence models. In **Interspeech**, 2016.

[39] Anuroop Sriram, et al. Cold fusion: Training seq2seq models together with language models. In **Interspeech**, 2017.

[40] Tara N. Sainath, et al. Two-pass end-to-end speech recognition. **ArXiv abs/1908.10992**, 2019.

[41] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **ArXiv**, Vol. abs/2106.09685, , 2021.