The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

The NTNU ASR System for Formosa Speech Recognition Challenge 2023

Hao-Chien Lu, Chung-Chun Wang, Jhen-Ke Lin, Tien-Hong Lo
Speech and Machine Intelligence Laboratory, National Taiwan Normal University
{howchien,takala,jacob,teinhonglo}@ntnu.edu.tw

摘要

近年，隨著大型語音處理模型如 Whisper 的快速發展，高辨識率的自動語音辨識 (Aautomatic Speech Recognition, ASR) 系統已成可能。儘管 Whisper 在主要語言上的表現卓越，非主要語言如臺語和客語的辨識率仍需提升。政府推動的「國家客家發展計畫」強調客家文化的重要性，因此開發客語 ASR 系統具有重要性。本文提出了一個客語 ASR 系統，參加了 2023 年的福爾摩沙語音辨識競賽 (FSR-2023)。我們使用了 Whisper 模型，經由 680,000 小時的語音-文本訓練，並運用遷移學習和 LoRA 技術使其適用於客語。實驗結果顯示，我們的方法在客語拼音和客語漢字的辨識上均取得了優異成績。未來，我們計劃繼續優化模型，使其適用於更多的台灣語言。

Abstract

In recent years, with the rapid development of large-scale speech processing models like Whisper, high-recognition ASR systems have become achievable. Although Whisper performs excellently in major languages, the recognition rates for non-major languages such as Taiwanese and Hakka still need improvement. The government's "National Hakka Development Plan" emphasizes the significance of Hakka culture, making the development of a Hakka ASR system paramount. This paper presents a Hakka ASR system, participating in the Formosa Speech Recognition Competition 2023(FSR-2023). We utilized the Whisper model, trained with 680,000 hours of speech-text data, and applied transfer learning and LoRA techniques to adapt it for Hakka. Experimental results demonstrate that our approach achieved commendable results in both Hakka phonetic and character recognition. In the future, we plan to further optimize the model to make it applicable to more Taiwanese languages.

關鍵字：福爾摩沙語音辨識競賽、Whisper、LoRA

Keywords: Formosa Speech Recognition Challenge, Whisper, LoRA

1 簡介

近年來，隨著 Whisper 大型語音處理模型快速發展，我們已經能輕鬆做出高辨識率的自動語音辨識系統。然而，由於 Whisper 使用網路上收集的語音-文本 (Speech-Text) 資料對上進行訓練，因此在主要語言 (如英語、國語) 上的辨識率較高，非主要語言 (如臺語、客語) 則相對較低。

政府在近年推動「國家客家發展計畫」，代表客家文化的重要地位。除了保存客家文化外，客語在日常生活的應用，如語音助理、自動生成字幕等，成爲了待發展的項目。

本文介紹由我們提出的客語 ASR 系統，用於 2023 年福爾摩沙語音辨識競賽 (FSR 2023)。客語的發音非常多樣，相較英文發音而言挑戰性更高。現有的大型語音處理模型，如 Whisper，經過 680,000 小時語音-文本 (Speech-Text) 的訓練，因此我們使用遷移學習技術，讓 Whisper 可以運用在客語上。雖然 FSR-2023 允許參賽者使用外部的客語資料進行訓練，但基於客語語料較少，我們使用少量的客語語料進行遷移學習，並使用 Low-Rank Adaptation(LoRA) 進行微調。也探索了多種後處理技術。最終，我們分別在客語拼音項目與客語漢字項目中獲得第 2 名與第 3 名的成績。

2 打造客語語音辨識 (ASR) 模型的策略

本段說明我們如何做出客語語音辨識模型。我們的架構主要分爲兩部分，第一部分爲語音模型，第二部分爲後處理。

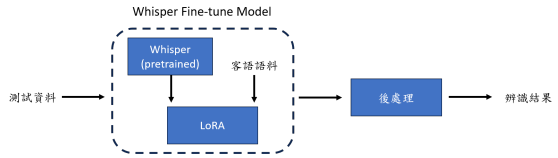

圖 1: 我們的客語語音辨識 (ASR) 模型架構

2.1 Branchformer

Branchformer 是由Peng et al. (2022) 提出的創新的語音識別架構。該模型的特點在充分捕獲語音資料中的區域 (local) 和全域 (global) 上下文。儘管過去的方法，如 Conformer，已經成功地結合了卷積 (convolution) 和自注意力 (self-attention) 來捕獲這些依賴關係，但 Branchformer 提供了一種更靈活、可解釋和可定制的編碼器替代方案。

Branchformer 的核心思想是利用平行分支 (parallel branches) 捕捉不同範圍的依賴關係。在每個編碼器層中，一個分支專注於捕獲長距離的依賴關係，而另一個分支使用多層感知機 (Multi-Layer Perceptron, MLP) 來捕捉本地依賴關係。此外，模型採用了一個名爲 gating Multi-Layer Perceptron (gMLP) 的門控 MLP 版本，該版本在先前的視覺和語言任務中都有不錯的表現。

通過廣泛的實驗，Branchformer 在多個自動語音識別和語言理解 (Spoken Language Understanding, SLU) 基準測試中均表現出色，超越了當時最先進的方法，如 Transformer 和 Convolutional Gated MLP (cgMLP)(Rajagopal and Nirmala, 2021)。此外，它的兩分支設計使得模型可以輕鬆地進行修改。最後，模型的權重學習策略使其在不同的層中使用區域和全域依賴關係，這有助於更深入地設計和調整模型。

2.2 WavLM

WavLM (Waveform Language Model) 是由Chen et al. (2022) 提出的大規模自監督預訓練 (Self-Supervised Pre-Training, SSL) 語音處理模型。該模型在 94,000 小時的 LibriLight 語音資料上進行了訓練，專爲解決全堆疊語音處理的挑戰而設計。與 Wav2vec 2.0(Baevski et al., 2020) 和 HuBERT(Hsu et al., 2021) 等其他知名自監督語音模型相比，WavLM 在 SUPERB (Speech processing Universal PERformance Benchmark) 基準上展現出了顯著的改善，並在包括語音辨識等多個語音任務中實現了當時最先進的性能。

WavLM 的主要特點是其能夠捕捉語音序列的順序結構，並在預訓練時學習到深層的語音和文本訊息。此外，與過往模型比較，該模型還採用了更大的訓練資料集，涵蓋了多種語言和語境，從而進一步強化其泛化能力和多語言處理能力。它的結構也有所優化，以進一步提高訓練速度和模型效能。

2.3 Whisper

Whisper 是由 OpenAI 提出的大詞彙連續語音處理模型[1]。該模型在超過 680,000 小時的網路上收集的語音-文本資料對上進行了訓練，能夠完成多語言 (Multilingual) 的自動語音辨識和語音翻譯 (Machine Translation) 任務。與 Wav2vec 2.0 或 HuBERT 等自監督語音預訓練模型 (Self-supervised Speech Pre-trained Model) 相比，Whisper 在多個公開資料集上達到了當時最先進的性能，並對環境變異 (例如噪音和口音) 具有強健性 (Robustness)(Radford et al., 2022)。

Whisper 模型屬於常見的 Attention Encoder-Decoder Model，並使用了標準的 Transformer 架構 (Vaswani et al., 2017)。在 Transformer 編碼器之前，模型額外加入了兩層 1D 卷積層 (Convolution Neural Network)，用於對輸入的梅爾頻譜特徵 (Mel Spectrogram) 進行下採樣 (Downsample)。OpenAI 釋出的模型有多個參數量大小：最小的模型爲 4 層 Encoder 和 4 層 Decoder 的 tiny 版本，參數量爲 33M；最大的模型爲 32 層 Encoder 和 32 層 Decoder 的 large 版本，參數量爲 1550M。標記器 (Tokenizer) 採用 Byte Pair Encoding (BPE) 來標記英語模型 (包括 tiny, base, small, and medium) 以及多語言模型 (包括 tiny, base, small, medium, and large，其中 large 分爲 large-v1 和 large-v2，且 large-v2 的整體表現最佳)。模型表現數據和參數細節可參考以下網址[2]。

2.4 Parameter-efficiency fine-tune (PEFT): LoRA

遷移學習是一種常用的技術，用於調整預訓練模型以適應不同的任務或領域 (Wang et al., 2021)。然而，大多數任務的遷移學習通常需要向預訓練模型添加額外的參數，以應對不同任務間的顯著差異 (Yang et al., 2020; Wang et al., 2020)。相對而言，由於 Whisper 模型已經在語音到文本的任務上進行了預訓練，並且客語漢字標籤空間是 Whisper 詞彙空間的子集，因此我們可以直接使用序列到序列

[1] https://github.com/openai/whisper/blob/main/README.md
[2] https://github.com/openai/whisper/blob/main/README.md

(Sequence-to-Sequence) 的框架在客語漢字上進行微調，而無需添加額外參數。考慮到運算資源的限制，我們還嘗試了一種低成本的微調方法，即 LoRA(Hu et al., 2022)。LoRA 首先固定了原始預訓練模型的參數，然後在 Transformer 的注意力層的線性權重中注入兩個低秩 (Low-Rank) 分解矩陣：降維矩陣 (Downsample Matrix)W_{down} 和升維矩陣 (Upsample Matrix)W_{up}。這樣做大幅減少了在微調過程中需要更新的參數數量。

2.5 使用 LoRA 進行微調 (Fine-tuning)

在實驗設置中，我們針對 Whisper-large-v2 模型應用了 LoRA 技術。在所有的注意力層 (Attention Layers) 中，我們皆加入了 LoRA：W_{down} 和 W_{up} 的權重矩陣，其中設定 Rank 爲 8， 參數設定爲 16。所有的 ASR 模型都在一張 NVIDIA 3090 GPU 上進行了 10 個訓練週期 (epochs)。我們使用 AdamW 作爲 Optimizer，並將學習速率 (Learning Rate) 設定爲 $5.0e-04$。

2.6 後處理

2.6.1 語言模型 (Language Model)

Transformer-based Language Models 倚賴自注意力機制（self-attention mechanism）來處理序列數據，特別適用於處理文本數據。在 ASR 模型做解碼 (decode) 後，可能會有模糊或不確定的情況，此時可以利用語言模型，根據上下文來糾正錯誤識別的單詞。我們使用客家委員會授權的「臺灣客語語料庫」[3]資料訓練語言模型。

2.6.2 淺融合 (Shallow Fusion)

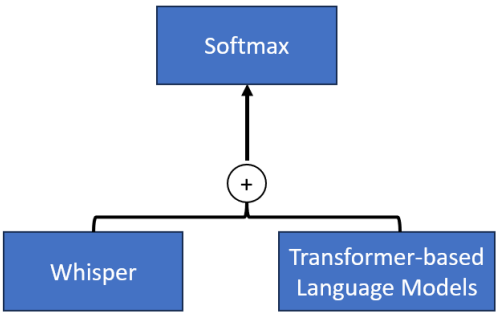

圖 2: 淺融合架構

淺融合 (Shallow Fusion) 是一種在解碼時將外部語言模型與 ASR 模型融合的方法。架構如圖 2。將外部語言模型訓練好的 Whisper 模型與訓練好的語言模型會輸出機率向量，將兩個機率向量加權加總後得到最終的機率向量。其中，可以利用 λ 調整語言模型的權重。我們將 λ 設爲 0.2。

2.6.3 重新計分 (N-best Rescoring)

重新計分方法將 ASR 輸出的前 n 個候選句利用語言模型重新評分，使得上下文會被考慮，提高候選句的合理性，最後重新產生 n 個候選句。我們使用 10-best 進行重新計分。

我們嘗試使用兩種技術進行重新計分，第一種爲單純使用語言模型，即使用 Transformer-based Language Models 進行重新計分。第二種我們使用了 pBERT。

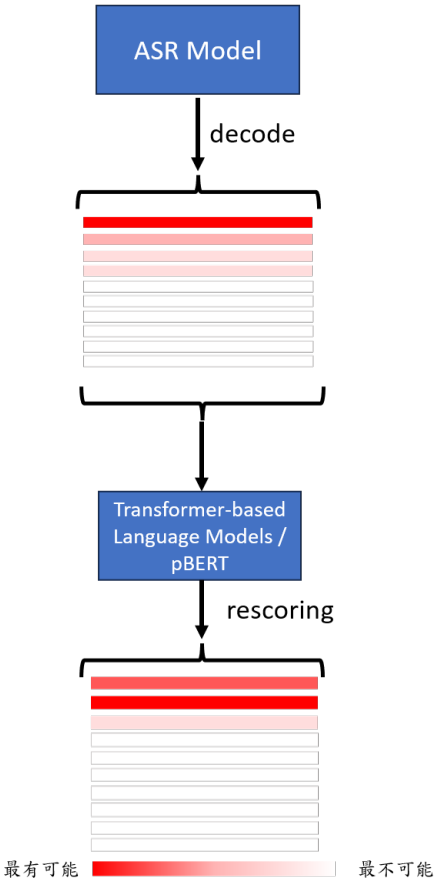

圖 3: 重新計分架構

此外，我們還嘗試使用 pBERT 作爲另一種重新計分的選項。pBERT 是在 BERT 的基礎上增加了一層全聯接層 (Fully Connected Layer, FC) 並通過微調一層附加的輸出層，希望能透過 BERT 來使其在重新計分上能得到更好的結果 (Chiu and Chen, 2021)。

3 實驗

3.1 資料集

我們使用 FSR-2023 競賽官方提供的訓練資料 集 FSR-2023-Hakka-Lavalier-Train(FSR-

[3] https://corpus.hakka.gov.tw/

表 1: 資料集配置

	Train			
	語者數	句數	字數	小時
train	60	16299	291042	47.45
dev	8	2126	38390	6.15
test	8	2187	39659	5.88
	Train3			
	語者數	句數	字數	小時
train	76	20612	369091	59.49
dev	11	3598	51025	10.01
test	11	3598	51025	10.01

Train)[4]及前測資料集 FSR-2023-Hakka-XYH8X-Eval(FSR-Eval)[5]。在訓練集 Train 中，我們使用 FSR-Train 作爲訓練、驗證與測試資料。在訓練集 Train3 中，我們使用 FSR-Train 作爲訓練資料，FSR-Eval 作爲驗證與測試資料。

3.2 後處理比較

我們使用淺融合、重新計分、淺融合 + 重新計分及 pBERT 等四種不同的後處理方法，比較經過後處理的性能。

在此實驗中，我們使用 Branchformer 作爲 ASR 模型，並且使用 Train 作爲訓練與測試資料。

在客語漢字中觀察到，使用所有後處理方法

表 2: 客語拼音後處理比較

客語拼音	
	WER(%)
無後處理	5.0
淺融合	5.1
重新計分	5.7
淺融合 + 重新計分	5.7
pBERT	5.0

皆比未使用後處理的效果還差。因此客語拼音上，我們不會使用後處理。

在客語漢字中觀察到，使用重新計分方法相較於其他後處理方法，有較佳的結果。這也和客語拼音不同。在客語拼音中的重新計分比無後處理效果還差，但在客語漢字中卻有較佳的結果。因此客語漢字上，我們使用重新計分方法作爲後處理的方法。在 pBERT 上，在客語

[4] https://speech.nchc.org.tw/FSR-2023/FSR-2023-Hakka-Lavalier-Train

[5] https://speech.nchc.org.tw/ntut/FSR-2023-Hakka-XYH8X-Eval

表 3: 客語漢字後處理比較

客語漢字	
	CER(%)
無後處理	4.5
淺融合	4.5
重新計分	4.2
淺融合 + 重新計分	4.2
pBERT	4.5

漢字與客語拼音皆未產生改變，因此我們最後也並未將 pBERT 作爲後處理的方法。

3.3 語音模型比較

我們使用不同的語音模型，比較在客語漢字和客語拼音上的效果。我們使用了兩種資料集進行訓練:

在客語拼音中，我們使用 Fbank(FBK)(Pariente et al., 2020)、Fbank+Pitch(FBK+Pitch)、WavLM 等三種提取聲音特徵的技術，並使用 Byte Pair Encoding(BPE)、Syllable(SYB) 等兩種代碼(Token) 化方法，以及使用 Conformer(CFR)、Branchformer(BFR) 等兩種語音模型進行比較。結果如圖 4。由結果得知使用 WavLM + SYB + CFR 在 Test 或是 Pilot_test 皆得到最低的 SER。

Train set	Feats	Token	Model	SER	
				Test	Pilot_test
Train	FBK	BPE	BFR	5.30	-
	FBK	SYB	BFR	5.25	-
	FBK+Pitch	BPE	BFR	5.95	-
	FBK+Pitch	SYB	BFR	4.96	-
	WavLM	BPE	CFR	4.03	6.21
	WavLM	SYB	CFR	3.89	6.31
	WavLM	SYB	BFR	3.90	6.39
Train3	WavLM	SYB	CFR	-	5.74
Train (Hanzi)	FBK	CHAR	WSP_LGv2 + MT	4.46	8.39

圖 4: 不同模型在客語拼音的 SER 比較

在客語漢字中，我們使用 FBK 做提取聲音特徵，使用 Character(CHAR)、WSP_ML 等兩種代碼化方法，以及使用 BFR、Whisper-large-v2(WSP_LGv2) 等兩種語音模型進行比較，其中 WSP_LGv2 使用 LoRA 進行微調。結果如圖 5。

我們發現在客語漢字語音模型中，WSP_LGv2 的效果最佳，因此我們使用這個模型作爲最終比賽的模型。雖然在 Branchformer 使用重新計分的結果較佳，但受限於比賽繳交結果時間，最後我們並沒有使用後處

Train set	Feats	Token	Model	CER	
				Test	Pilot_test
Train	FBK	CHAR	BFR	4.11	23.53
	FBK	WSP_ML (ZH)	WSP_LGv2	1.74	8.38
Train3	FBK	WSP_ML (ZH)	WSP_LGv2	0.19	7.09

圖 5: 不同模型在客語漢字的 CER 比較

理。

4 結果

4.1 熱身賽結果

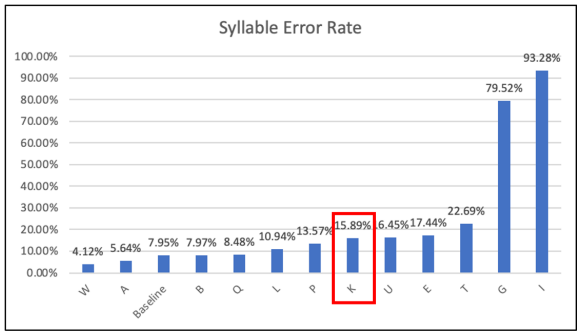

圖 6: 熱身賽客語拼音結果，紅框表示我們的成績

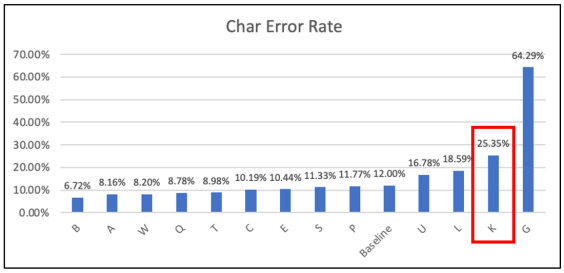

圖 7: 熱身賽客語漢字結果，紅框表示我們的成績

熱身賽我們使用大會提供的基礎模型，客語拼音使用 Conformer + WavLM，客語漢字使用 Transformer。我們在熱身賽僅有確認整體流程是否正確，並沒有使用優化過的模型。圖 6 及圖 7 爲熱身賽結果。

4.2 決賽結果

決賽我們在客語拼音使用 WavLM ，客語漢字爲 Whisper+LoRA。但由於繳交時間不足，我們並沒有做重新計分。圖 8 及圖 9 爲決賽結果。我們在客語拼音獲得第二名，客語漢字獲得第三名。

5 結論與未來發展

在本文中，我們介紹了參加 2023 福爾摩沙語音辨識競賽，通過一系列的實驗評估，我們使用了 Whisper+LoRA 作爲客語辨識模型，並

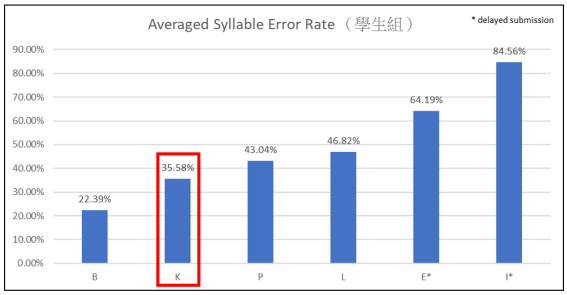

圖 8: 決賽客語拼音結果，紅框表示我們的成績

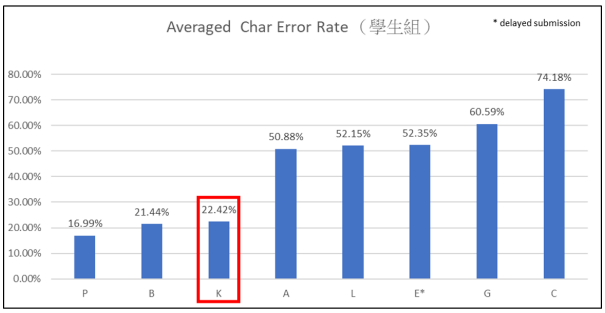

圖 9: 決賽客語漢字結果，紅框表示我們的成績

使用語言模型、淺融合、重新計分進行後處理。至於未來工作，我們計劃研究更高級的端到端方法，用於台灣語言聲學建模，並將我們的建模策略應用於不同少數語言的 ASR 任務。

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Shih-Hsuan Chiu and Berlin Chen. 2021. Innovative bert-based reranking language models for speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Filterbank design for end-to-end speech separation.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

A. Rajagopal and V. Nirmala. 2021. Convolutional gated mlp: Combining convolutions & gmlp.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Minghan Wang, Jiaxin Guo, Yimeng Chen, Chang Su, Min Zhang, Shimin Tao, and Hao Yang. 2021. Make the blind translator see the world: A novel transfer learning solution for multimodal machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 139–149, Virtual. Association for Machine Translation in the Americas.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 53–59, Lisboa, Portugal. European Association for Machine Translation.

Hao Yang, Minghan Wang, Ning Xie, Ying Qin, and Yao Deng. 2020. Efficient transfer learning for quality estimation with bottleneck adapter layer. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 29–34, Lisboa, Portugal. European Association for Machine Translation.