INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

WhiSLU: End-to-End Spoken Language Understanding with Whisper

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, Hao Yang

Huawei Translation Services Center

{wangminghan,liyinglu,guojiaxin1,qiaoxiaosong,lizongyao,shanghengchao,
weidaimeng,taoshimin,zhangmin186,yanghao30}@huawei.com

Abstract

Spoken Language Understanding (SLU) systems commonly use cascading structures. However, these systems are prone to error propagation, information loss, high costs, and latency, leading researchers to explore end-to-end (E2E) SLU as a hot topic. However, E2E SLU faces the challenge of insufficient data, resulting in most previous work relying on pretrained acoustic models. Nevertheless, pre-training task and SLU task solution spaces are often substantially different, making it difficult for E2E SLU models to surpass cascading models. To address this, we propose using OpenAI's Whisper model for SLU tasks. We employ the Sequence-level Multitask Learning (SML) paradigm, which encodes multiple ASR-related tasks into a sequence for learning. Our method significantly outperforms the E2E baseline by a large margin (with a 10% improvement in EM score) and even outperforms cascading models, achieving a 77% EM score on the STOP dataset, demonstrating its effectiveness.

Index Terms: Spoken Language Understanding, SLU, Transfer Learning, Multitask Learning

1. Introduction

Over the years, speech recognition has gained significant attention due to its increasing usage in popular customer devices such as Alexa, Siri, and more. As the use of these systems becomes more prevalent, accurately recognizing user intent and desires has become an important task. Traditionally, an SLU system has been implemented through a cascaded approach by concatenating a set of components. These components typically include an ASR model that transcribes spoken information into transcripts, and an NLU model that extracts intent and fills key entities into pre-defined slots to create formalized instructions [1].

However, cascaded systems have the possibility of propagating errors through components, which can lead to undesirable results. For example, the absence of a single keyword in an ASR transcript can lead to a completely different outcome by the NLU system. Additionally, cascaded systems have the risk of losing important information. The intonation and emphasis of speakers cannot be explicitly expressed in ASR transcripts.

To address these issues, E2E models are proposed that predict intent directly from given audio. This approach eliminates the need for concatenating multiple components and provides a more accurate and efficient way to recognize the user's intent.

Although the E2E approach offers several advantages, the limited amount of available data makes it difficult to train an E2E model from scratch. As a result, previous studies have primarily relied on transfer learning to adapt pretrained acoustic models, such as Wav2vec 2.0[2, 3] or Hubert [4], to the SLU task. However, the significant difference in the solution space between the pretraining task and the downstream task often results in difficulty with transfer learning. For example, while pretrained models aim to learn low-level acoustic representation at the pre-training stage, they require learning high-level semantic information in the challenging SLU task. Therefore, some approaches [5] have also employed pretrained ASR models, but the scale of the model and pretraining dataset remains limited. Consequently, none of these approaches have outperformed the cascade method.

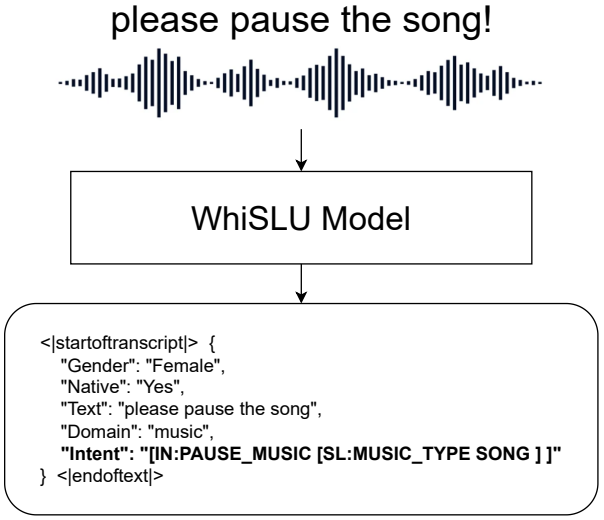

Figure 1: *This figure presents an example of WhiSLU's prediction. The model is trained to directly generate a well-formatted JSON string, learned from the sequence-level multitask learning strategy. The "Intent" entry represents the main task prediction, while the remaining entries correspond to the predictions of auxiliary tasks.*

In this paper, we propose a novel approach to tackle these challenges and enhance the efficacy of using pretrained models for spoken language understanding. Specifically, we explore two research questions: (1) How to alleviate the difficulty of task transfer learning when there are substantial differences in solution spaces, and (2) how to enhance the efficiency of fine-tuning large-scale ASR models.

To address these research questions, we use a pretrained model called Whisper[6], which integrates multilingual Automatic Speech Recognition (ASR) and speech translation tasks. The original tasks of Whisper have significant overlap with the downstream task of SLU, allowing us to leverage the model for

10.21437/Interspeech.2023-1505

Intent	Percentage of Intent	Slot	Percentage of Slot	Gender	Percentage of Gender	Native English	Percentage of whether native speaker
GET_WEATHER	14.60%	**DATA_TIME**	23.93%	**Female**	44.55%	**Native**	95.18%
CREATE_ALARM	7.48%	**LOCATION**	7.89%	**Male**	54.62%	**Non-native**	4.09%
CREATE_REMINDER	7.03%	**TODO**	7.25%	**Non-binary**	0.79%	**Prefer not to say**	0.73%
...	...	...	...	**Unanswered**	0.04%	-	-

Table 1: *STOP statistics about intent, slot, gender and native English speakers. Specifically, the dataset comprises 80 distinct intents and 82 different slots, with only the top 3 being displayed. The gender distribution is composed of 4 classes, wherein both the male and female categories constitute approximately half of the total number of recordings. Regarding the proficiency of English speakers, the majority of audio files belong to native speakers, as evidenced by the proportion of native speakers being higher than non-native speakers.*

file_id	domain	gender	native	utterance	decoupled_normalized_seqlogical
eval_0/alarm_eval_0 /00002321.wav	alarm	Female	Yes	*Set an alarm tomorrow*	[IN:CREATE_ALARM [SL:DATE_TIME]]
eval_1/alarm_eval /00000938.wav	alarm	Female	Yes	*where are my alarms?*	[IN:GET_ALARM]

Table 2: *STOP dataset semantic parse samples.*

SLU with minimal additional training.

To further enhance the effectiveness of our approach, we propose a Sequence-level Multitask Learning paradigm, which utilizes multiple ASR-related tasks such as domain classification, speaker gender classification, and accent classification to extract critical information level by level for the final NLU task. The labels of these tasks are concatenated into a text sequence with the NLU labels as the last part, which can be trained through the auto-regressive text generation paradigm, and thereby requires no additional parameters.

This paper makes the following contributions:

- We introduce WhiSLU, an end-to-end SLU model that leverages sequence-level multitask learning to perform transfer learning with the pretrained Whisper model.
- Through our experiments evaluated on the STOP [5] dataset, we demonstrate that WhiSLU significantly outperforms both end-to-end (by 10%) and cascade (by 5%) SLU models, highlighting its superior performance and efficiency.

2. Method

2.1. Background

2.1.1. SLU Dataset

Spoken Task-Oriented Semantic Parsing (STOP) [5] is a recently released SLU dataset built on top of the previous text-only NLU dataset TOPv2 [7]. All utterances in STOP are recorded by humans and undergo rigorous quality checking. In addition to surpassing previous datasets, such as FSC [8] and SLURP [9], in terms of data size, STOP also poses a greater challenge due to its higher complexity, containing numerous semantic parses with multiple layers of nesting, and a greater number of intents and slots. Furthermore, STOP provides additional low-resource splits and synthetic speech generated through TTS to support low-resource SLU research. In our paper, our model is tested mainly on the high-resource split. Some of the examples are shown in Table 2. Table 1 presents the statistics of the STOP dataset.

2.1.2. Whisper

Whisper [6] is a large-scale speech processing model proposed by OpenAI. It was trained on a collection of over 680,000 hours of speech-text pairs collected from the internet and is capable of performing multilingual ASR and speech translation tasks. Whisper can handle not only short audio segments but also long audio thanks to its built-in time labeling mechanism.

The Whisper model uses the standard Transformer [10] architecture with an additional two layers of 1D convolution set before the encoder for down-sampling the input mel-spectrogram features. Whisper achieves state-of-the-art performance on multiple publicly available datasets and has strong robustness. Compared to other unsupervised speech pretraining models such as Wav2vec 2.0 [2] or Hubert [4], Whisper's output space is usually closer to downstream tasks, making it more suitable as a pretrained model for task transferring. Meanwhile, the scale of the training set is also larger than the previous works. Therefore, we chose Whisper as the pretrained model for SLU.

2.2. Efficient Transfer Learning

Transfer learning is a commonly used technique to adapt a pretrained model for a different task or domain. However, most task-level transfer learning requires adding additional parameters to the pre-trained model to deal with the significant differences between both tasks [11]. For example, when fine-tuning a pretrained BERT [12] model on sequence labeling tasks, an additional classifier is typically needed [13, 14]. Similarly, using pretrained acoustic models like Wav2vec 2.0 in the SLU task often requires adding an extra decoder. In contrast, since the Whisper model is already pretrained on the speech-to-text task and the SLU label space is a subset of the Whisper vocabulary space, we can directly fine-tune Whisper on SLU labels using the sequence-to-sequence paradigm without additional parameters.

In addition to performing full-parameter fine-tuning on different sizes of Whisper models (except for large, where we froze the encoder), we also attempted a lower-cost fine-tuning approach, i.e. LoRA [15]. LoRA first fixes the parameters of the original pretrained model and injects two low-rank decomposition matrices into the linear weights of attention layers in the

Transformer, greatly reducing the number of parameters that need to be updated during fine-tuning. In the experiment, we only used LoRA for the large model.

2.3. Sequence Level Multitask Learning

Although Whisper performs well on the ASR task, the model does not necessarily require strong semantic understanding capabilities because we found that training the model solely on NLU labels using seq2seq training can easily lead to a performance bottleneck. Additionally, we found that there are often many errors in the leaf nodes of the SLU semantic parse tree predicted by the model, which may be due to the model losing its original ASR capabilities during SLU training, a phenomenon known as catastrophic forgetting. Therefore, we propose a method called Sequence Level Multitask Learning (SML), which combines labels of multiple tasks into a sequence for the model to learn through seq2seq training. These tasks include speaker gender classification, speaker nativeness classification, ASR, domain classification, and SLU. These tasks are arranged in the above order to extract information from acoustics to semantics step by step, assisting the final SLU task.

Formally, we denote the set of related tasks as $\mathcal{T}$, the label space of each task as $\mathcal{Y}_t$, where $t \in \mathcal{T}$, and the vocabulary of the Whisper model as $\mathcal{V}$. Using the Whisper tokenizer ϕ, we can tokenize the labels of each task into a sequence of sub-words (tokens) that belong to the vocabulary of Whisper: $\mathbf{y}_t^{\mathcal{V}} = \phi(y_t)$, where $y_t \in \mathcal{Y}_t$. This enables us to convert task-specific labels from their respective label spaces into the vocabulary space of Whisper, allowing us to formulate them as a sequence-to-sequence task:

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{t \in \mathcal{T}} P(\mathbf{y}_t^{\mathcal{V}} | \mathbf{y}_{<t}^{\mathcal{V}}, \mathbf{X}; \theta), \tag{1}$$

where $\mathbf{y}_{<t}^{\mathcal{V}}$ represents all tokens before the current task, and θ is the parameter set of the Whisper model or possibly introduced parameters.

During our experiment, we encoded all labels into a JSON string (as illustrated in Figure 1) and tokenized it with the Whisper tokenizer, instead of simply concatenating label tokens. Our findings show that the encoded label sequence with an organized format and explicit keys as separators can aid the model in learning multiple tasks with greater ease. Additionally, this approach also simplifies the post-processing.

3. Experiments

3.1. Experimental Setup

We evaluated our approach using the Whisper-base, Whisper-medium, and Whisper-large models. For the base and medium models, we fine-tuned all parameters, while for the large model, we froze the encoder and only fine-tuned the parameters in the decoder. We only apply LoRA on the Whisper-large model with the W_Q and W_V weight matrices in all attention layers being injected, the rank r is set as 32 and α set as 64. We trained all models with 4 V100 GPUs for 20,000 steps, setting the learning rate to 1e-4 and the batch size equivalent to 8 per card with accumulation steps being varied for each model.

3.2. Metrics

Our primary evaluation metric is Exact Match accuracy (EM), which assesses the quality of the predicted SLU sequence by assigning a value of 1 if all tokens match the ground truth sequence and 0 otherwise. The resulting score is then normalized by the size of the testing set.

To evaluate the structure accuracy of the semantic parsing tree, we also use Tree EM, as suggested in [5], which ignores leaf nodes in the predicted slots during matching. Additionally, we use WER as a more tolerant metric.

All semantic-parse related metrics are computed based on the text in the *decoupled_normalized_seqlogical* field of the original manifest file.

Furthermore, we report the WER score of the predicted transcripts, as well as the Mathews Coefficient (MCC) of the auxiliary classification tasks.

3.3. Baseline Systems

We compare WhiSLU against four types of baseline systems:

- The cascade system proposed in [5], which consists of an ASR model (pretrained Wav2vec 2.0 [2] finetuned on the STOP training set) and an NLU model (pretrained BART-base [16] finetuned on the STOP text data).
- A pretrained Wav2Vec 2.0 [2] model with an additional attention decoder. It is first finetuned it on the STOP ASR transcripts and then on the STOP semantic parses [5].
- A pretrained Hubert [4] model with an additional attention decoder. It is first finetuned it on the STOP ASR transcripts and then on the STOP semantic parses [5].
- The Whisper model with three sizes, directly finetuned on STOP semantic parse sequences.

3.4. Experimental Results

3.4.1. Overall Comparison

Table 4 presents the results of WhiSLU on the STOP test set. Compared with the baseline models provided by the official baseline, WhiSLU significantly outperforms them, regardless of whether SML is used or not, and also surpasses the cascade system with a large margin. This demonstrates that for SLU task transfer, models like Whisper that are directly pretrained on the speech-to-text task have stronger advantages than models that undergo unsupervised pretraining on acoustic features and then finetuning on limited ASR data.

3.4.2. Effectiveness of Sequence Level Multitask Learning

Table 3 presents the results of Whisper models of different sizes trained with or without Sequence Level Multitask Learning (SML). It is clear that using SML leads to consistent performance improvements (1-3 points in EM), fully demonstrating the effectiveness of SML. From a model size perspective, larger models can achieve better performance in both settings, which is also in line with expectations. Besides, we further analyzed which auxiliary tasks are more helpful in improving the performance of the SLU task and found that ASR task provides the greatest help to SLU, followed by domain classification, while gender and nativeness have little impact on SLU. An unexpected result was observed in the classification performance of gender and nativeness, where WhiSLU-large-SML did not achieve the best performance. This could be attributed to the frozen encoder parameters, which may not have fully adapted to the acoustic features of the STOP dataset. Since both tasks are more related to low-level acoustic information extraction, this could have contributed to the suboptimal performance.

Model	EM	EM-TREE	ASR-WER	MCC-domain	MCC-gender	MCC-nativeness
WhiSLU-base	68.32	81.57	-	-	-	-
WhiSLU-base-SML	71.26	83.03	16.18	0.987	0.9477	0.3537
WhiSLU-medium	73.3	85.17	-	-	-	-
WhiSLU-medium-SML	74.13	85.46	10.85	0.9874	0.9463	0.4054
WhiSLU-large	74.49	84.89	-	-	-	-
WhiSLU-large-SML	**76.68**	**86.37**	**3.19**	0.9889	0.9017	0.3572

Table 3: *Comparison of seq-level multitask learning. We conducted a comparison of the performance of the Whisper model at different sizes (base, medium, and large) with and without sequence-level multitask learning (SML). The results indicate that employing SML can lead to improved EM and EM-TREE performance, across all three model sizes. Specifically, we observed a positive correlation between EM, EM-TREE, and ASR-WER for the ASR prediction subtask. For the other three subtasks, which involve the prediction of Domain, Gender, and Nativeness, we used the Matthews Correlation Coefficient (MCC) to evaluate the performance of the WhiSLU models, and found only small differences among the three different sizes.*

Model	EM	EM-TREE	SLU-wer
wav2vec2.0 [5]	68.70	82.78	-
HuBERT [5]	69.23	82.87	-
cascade system	72.36	82.78	-
WhiSLU-large	74.49	84.89	6.8103
WhiSLU-large-SML	**76.68**	**86.37**	**6.1407**

Table 4: *Overall comparison result. The table illustrates a comparison between the performance of the baseline models and our proposed WhiSLU models. Notably, both the WhiSLU-large models with and without the sequence-level multitask learning (SML) strategy achieved significant improvements in EM, EM-TREE, and SLU-wer. This suggests that our proposed method of transferring knowledge from the ASR model to the NLU task is effective in enhancing the performance of the WhiSLU model.*

Model	# Trainable Params	EM	EM-TREE
WhiSLU-base-SML	74M	68.32	81.57
WhiSLU-large-LoRA	15M	72.96	83.03
WhiSLU-large-LoRA-SML	15M	71.98	81.33
WhiSLU-large-SML	1550M	**76.68**	**86.37**

Table 5: *Finetuning Efficiency with LoRA. The findings demonstrate that utilizing LoRA on the WhiSLU-large model requires training only 1% of the parameters to outperform the performance achieved through full parameter finetuning with the WhiSLU-base model. This implies a substantial enhancement in finetuning efficiency.*

3.4.3. Finetuning Efficiency with LoRA

Table 5 displays the results of WhiSLU after finetuning with LoRA. The outcomes illustrate that using LoRA on Whisper-large only necessitates training 1% of the parameters to surpass the performance of full parameter finetuning with Whisper-base, resulting in a significant improvement in finetuning efficiency. However, we discovered that using LoRA and SML concurrently can result in the failure of SML training. Upon analysis, we found that LoRA is generally useful for domain transfer [15], but it faces difficulties with task transfer, especially in multi-task finetuning. In this scenario, the model's performance on the primary task may decrease, and may not be as good as using the primary task as the sole target task (WhiSLU-large-LoRA is better than WhiSLU-large-LoRA-SML).

4. Conclusion

In this paper, we aimed to apply the transfer learning approach to the Whisper model for the SLU task and proposed the WhiSLU model. In contrast to traditional sequence-to-sequence finetuning, we introduced the sequence-level multitask learning paradigm, which prioritizes tasks according to their semantic complexity and concatenates their labels into a formatted JSON sequence for the model to learn to generate directly. This paradigm allows for smoother task transfer learning and improves the main task's performance by leveraging auxiliary task predictions. Additionally, we explored using LoRA for efficient finetuning of models with large parameter sizes and achieved strong baseline performance by training only 1% of the parameters. Finally, experimental results demonstrated that WhiSLU significantly outperformed E2E and cascade baselines on the STOP dataset, achieving state-of-the-art performance. Our future work will focus on enhancing WhiSLU's performance in low-resource scenarios.

5. References

[1] L. Qin, T. Xie, W. Che, and T. Liu, "A survey on spoken language understanding: Recent advances and new frontiers," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Z. Zhou, Ed. ijcai.org, 2021, pp. 4577–4584. [Online]. Available: https://doi.org/10.24963/ijcai.2021/622

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 3465–3469. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1873

[4] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[5] P. Tomasello, A. Shrivastava, D. Lazar, P. Hsu, D. Le, A. Sagar, A. Elkahky, J. Copet, W. Hsu, Y. Adi, R. Algayres, T. A. Nguyen, E. Dupoux, L. Zettlemoyer, and A. Mohamed, "Stop: A dataset for spoken task oriented semantic parsing," in *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*. IEEE, 2022, pp. 991–998. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10022703

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[7] X. Chen, A. Ghoshal, Y. Mehdad, L. Zettlemoyer, and S. Gupta, "Low-resource domain adaptation for compositional task-oriented semantic parsing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 5090–5100. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.413

[8] Y. Qian, X. Bian, Y. Shi, N. Kanda, L. Shen, Z. Xiao, and M. Zeng, "Speech-language pre-training for end-to-end spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 7458–7462. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9414900

[9] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 7252–7262. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.588

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[11] M. Wang, J. Guo, Y. Chen, C. Su, M. Zhang, S. Tao, and H. Yang, "Make the blind translator see the world: A novel transfer learning solution for multimodal machine translation," in *Proceedings of the 18th Biennial Machine Translation Summit - Volume 1: Research Track, MTSummit 2021 Virtual, August 16-20, 2021*, K. Duh, F. Guzmán, and S. Richardson, Eds. Association for Machine Translation in the Americas, 2021, pp. 139–149. [Online]. Available: https://aclanthology.org/2021.mtsummit-research.12

[12] J. Devlin, M. Chang, K. Lee, and K. 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-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[13] H. Yang, M. Wang, N. Xie, Y. Qin, and Y. Deng, "Efficient transfer learning for quality estimation with bottleneck adapter layer," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, M. L. Forcada, A. Martins, H. Moniz, M. Turchi, A. Bisazza, J. Moorkens, A. G. Arenas, M. Nurminen, L. Marg, S. Fumega, B. Martins, F. Batista, L. Coheur, C. P. Escartín, and I. Trancoso, Eds. European Association for Machine Translation, 2020, pp. 29–34. [Online]. Available: https://aclanthology.org/2020.eamt-1.4/

[14] M. Wang, H. Yang, Y. Qin, S. Sun, and Y. Deng, "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, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, M. L. Forcada, A. Martins, H. Moniz, M. Turchi, A. Bisazza, J. Moorkens, A. G. Arenas, M. Nurminen, L. Marg, S. Fumega, B. Martins, F. Batista, L. Coheur, C. P. Escartín, and I. Trancoso, Eds. European Association for Machine Translation, 2020, pp. 53–59. [Online]. Available: https://aclanthology.org/2020.eamt-1.7/

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703