MoE-SLU: Towards ASR-Robust Spoken Language Understanding via Mixture-of-Experts

Xuxin Cheng[†], **Zhihong Zhu**[†], **Xianwei Zhuang**,
Zhanpeng Chen, Zhiqi Huang, Yuexian Zou[*]
School of ECE, Peking University, China
{chengxx, zhihongzhu, xwzhuang, troychen927}@stu.pku.edu.cn
{zhiqihuang, zouyx}@pku.edu.cn

Abstract

As a crucial task in the task-oriented dialogue systems, spoken language understanding (SLU) has garnered increasing attention. However, errors from automatic speech recognition (ASR) often hinder the performance of understanding. To tackle this problem, we propose MoE-SLU, an ASR-Robust SLU framework based on the mixture-of-experts technique. Specifically, we first introduce three strategies to generate additional transcripts from clean transcripts. Then, we employ the mixture-of-experts technique to weigh the representations of the generated transcripts, ASR transcripts, and the corresponding clean manual transcripts. Additionally, we also regularize the weighted average of predictions and the predictions of ASR transcripts by minimizing the Jensen-Shannon Divergence (JSD) between these two output distributions. Experiment results on three benchmark SLU datasets demonstrate that our MoE-SLU achieves state-of-the-art performance. Further model analysis also verifies the superiority of our method.

1 Introduction

Spoken Language Understanding (SLU) is a fundamental task in recent task-oriented dialogue systems, aimed at capturing the comprehensive semantics of human speech. It plays an important role in personal assistants like Amazon's Alexa, Apple's Siri, and Microsoft's Cortana (Young et al., 2013; Cheng et al., 2023b; Zhuang et al., 2024). SLU focuses on two typical subtasks: intent detection and slot filling (Tur and De Mori, 2011; Cheng et al., 2024). Intent detection can be regarded as a semantic classification task (Xu et al., 2021), aiming to predict the user's intent (Chen et al., 2022b; Zhou et al., 2022). Slot filling could be approached as a sequence labeling task, where the goal is to predict the slot for each token (Zhou et al., 2021; Zhu et al., 2024; Zhao et al., 2024; Song et al., 2024).

In the realm of SLU, there are two common approaches: pipeline methods and end-to-end methods. Pipeline methods involve the cascaded combination of automatic speech recognition (ASR) and natural language understanding (NLU). The ASR system transcribes acoustic input into text, which is then fed into the NLU component to tackle the specific task. On the other hand, end-to-end SLU methods directly generate predicted results without explicit separation of ASR and NLU (Huang et al., 2022; Seo et al., 2022; Dong et al., 2023c).

When it comes to pipeline SLU methods, they provide the advantage of seamless integration of external datasets and the utilization of pre-trained language models. However, they are prone to error propagation, where mistakes from the ASR system could adversely affect the accuracy of subsequent NLU processing. As a result, enhancing the ASR robustness of the SLU model becomes crucial.

An effective approach to mitigate the detrimental impact of errors arising from ASR is to learn error-robust representations for SLU. D'Haro and Banchs (2016) proposes a phrase-based machine translation system trained with the words and phonetic encoding to automatically correct the ASR results. Mani et al. (2020) proposes a machine translation model that learns a mapping from the out-of-domain ASR errors to in-domain terms found in corresponding reference files. Leng et al. (2021) utilizes edit distance-based alignments between the encoder and decoder to perform ASR error correction. Dutta et al. (2022) bootstraps the BART-based sequence-to-sequence model and leverages several phonetically grounded fine-tuning strategies to enhance the correction of errors in ASR predictions.

In this paper, we propose MoE-SLU, an ASR-Robust SLU framework based on the mixture-of-experts technique to make better use of the clean manual transcripts and the ASR transcripts. We propose three strategies to generate more transcripts from clean transcripts, which could simulate more

[†] Equal contribution.
[*] Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 14868–14879

kinds of errors in real scenarios. Then we utilize the mixture-of-experts method to weighted average the representations of generated transcripts, ASR transcripts, and the associated clean manual transcripts. By learning the weights of different transcripts in mixture-of-experts, the clean manual and ASR transcripts are handled differently. Both the weighted average of representations and the representations of ASR transcripts are used to calculate the cross entropy with the corresponding labels. In addition, we also regularize these two predictions via minimizing Jensen-Shannon Divergence (JSD) between the two output distributions. Experimental results demonstrate that our MoE-SLU outperforms previous ASR-Robust SLU models, and model analysis also verifies the advantages of our method. To sum up, the contributions of our work are three-fold:

- We propose an ASR-Robust SLU framework MoE-SLU, which uses the mixture-of-experts technique to make better use of the ASR transcripts and the clean manual transcripts.
- Experiments on three public SLU datasets display that MoE-SLU achieves new state-of-the-art performance, surpassing previous works.
- Model analysis further verifies that MoE-SLU can indeed improve ASR robustness more effectively than previous works.

2 Related Work

ASR-Robust Spoken Language Understanding SLU aims to understand the user's current goal via constructing semantic frames (Cheng et al., 2023c; Dong et al., 2022, 2023a). Since SLU usually faces the challenge of error propagation from the ASR system, there are increasing attempts being made to enhance ASR robustness in SLU. Various traditional methods have been proposed to solve SLU, including support vector machine (SVM) and recurrent neural network (RNN) (Haffner et al., 2003). Xu and Sarikaya (2013b) attempts to leverage log-linear models to achieve intent detection.

Recently, many classification approaches based on the neural network such as convolutional neural networks (CNN) (Xu and Sarikaya, 2013a; Ravuri and Stolcke, 2015) have been investigated. For instance, Xia et al. (2018) leverages a capsule-based neural network with self-attention for SLU. With the recent remarkable performance demonstrated by pre-trained models across different tasks (Zhu et al., 2023a,b; Feng et al., 2023; Wu et al., 2024; Huang et al., 2024b; Shen et al., 2024; Chen et al., 2024), researchers began to explore the application of BERT-based pre-trained models (Devlin et al., 2019) in the field of SLU. Huang et al. (2022) applies LAS (Chan et al., 2016) and BART (Lewis et al., 2020) as the pre-trained models and proposes a multi-task learning framework termed MTL-SLT. In this work, we utilize RoBERTa (Liu et al., 2019) and Data2vec (Baevski et al., 2022) to acquire invariant representations between clean manual transcripts and erroneous ASR transcripts.

Mixture-of-Experts Mixture-of-Experts is proposed by Jacobs et al. (1991); Jordan and Jacobs (1994), which allows for the processing of different examples utilizing independent expert modules and has been adopted in different domains. Hochreiter and Schmidhuber (1997) utilizes mixture-of-experts to construct the large-scale language models. With the development of attention (Cao et al., 2021; Zhu et al., 2022; Zhuang et al., 2022; Li et al., 2022a; Xin et al., 2022, 2023b; Xin and Zou, 2023; Yin et al., 2023; Xin et al., 2023a), researchers began to use mixture-of-experts to feed-forward networks to enhance the performance of Transformers. Shazeer et al. (2017) uses mixture-of-experts in sequence learning. He et al. (2018); Cho et al. (2019) uses mixture-of-experts in the generation tasks, and Peng et al. (2020) leverages mixture-of-experts to improve the performance of the translation models. In addition, Zhang and Feng (2021) proposes the mixture-of-experts wait-k policy to build a universal simultaneous machine translation method capable of delivering the high-quality translations with arbitrary latency. In our study, we apply mixture-of-experts technique to weighted average the representations of the generated transcripts, the ASR transcripts, and the associated clean manual transcripts, which allows the model to leverage the difference between the clean manual transcripts and the corresponding ASR transcripts more efficiently.

3 Method

In this section, we first propose three strategies to generate additional transcripts from the clean transcripts (§3.1). Then, we begin to introduce our proposed ASR-Robust SLU framework MoE-SLU, including a self-supervised contrastive learning module in pre-training (§3.2) and a mixture-of-experts module in fine-tuning (§3.3). Finally, we introduce the training objective during the pre-training stage and the fine-tuning stage (§3.4).

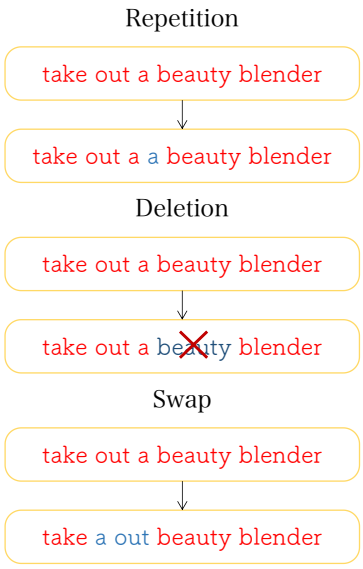

Figure 1: The illustration of the proposed strategies.

3.1 Proposed Data Augmentation Strategies

To capture a wider range of the linguistic variations and noise patterns in the real-world environments, we propose three strategies to generate some additional transcripts from the clean transcripts, thereby increasing the diversity of training data. The proposed strategies include repetition, deletion, and swap. Specific schematic illustrations of each operation are depicted in Figure 1 for better clarity. Since the changes brought about by these strategies are minor, they usually do not alter the underlying semantics of the original ASR transcripts. Moreover, in general, there are certain keywords in the transcript which play a crucial role in determining the intent. With the utilization of these three strategies, the ability of the SLU model to capture these keywords can be further improved.

Repetition Strategy As illustrated in the upper part of Figure 1, the repetition strategy involves randomly duplicating certain words in original ASR transcripts. In real environments, it is common for certain words to be repeated in the original speech. The corresponding intent label remains unchanged and the slots corresponding to the repeated words are added to the original slot labels. By incorporating the strategy into the data generation process, we can generate more transcripts that closely resemble the real-life scenarios to enhance the ability.

Deletion Strategy As illustrated in the middle part of Figure 1, similar to the repetition strategy, we can also delete some words which are relatively unimportant to generate some new transcripts. To achieve this, we utilize CoreNLP[1] (Manning et al., 2014) to extract non-nouns from the original transcripts. Subsequently, we randomly delete some of these non-nouns to generate new transcripts. The corresponding intent label also remains unchanged and the slots of the deleted words are deleted.

Swap Strategy As illustrated in the lower part of Figure 1, we also propose a swap strategy that randomly swaps two words in the original transcript to obtain the new transcript. The corresponding intent label remains unchanged and the slots of swapped words are swapped accordingly.

3.2 Self-supervised Contrastive Learning

Following previous works (Chang and Chen, 2022), we employ self-supervised contrastive learning during pre-training to develop sentence representations that are robust against to misrecognition and capable of handling the ASR errors. For a fair comparison, we utilize a pre-trained RoBERTa model (Liu et al., 2019) and Data2vec (Baevski et al., 2022). This continuous training process enables the SLU model to learn from the rich patterns and structures present in the input spoken language.

The mini-batch of input data is denoted as $B = (x^p, x^q)$, where x^p represents a clean manual transcript and x^q represents its associated ASR transcript. We first apply the proposed three strategies to generate additional transcripts from x^p. Then, we employ the pre-trained model and use the last layer representation of the special token `[CLS]` to obtain the corresponding representation h. By doing so, the model can capture the contextual information and encode it to meaningful representation.

We leverage the self-supervised contrastive loss $\mathcal{L}_{sc}$ (Gao et al., 2021) to adjust sentence representations. $\mathcal{L}_{sc}$ encourages similar representations for semantically related pairs of transcripts while pushing apart the representations of the unrelated pairs. This process could improve the ability of the model to capture underlying semantic information in the transcripts and promote the robustness against ASR errors (Wang and Isola, 2020):

$$\begin{aligned}\mathcal{L}_{sc} &= -\sum_{(h,h^+)\in P} \log \frac{e^{s(h,h^+)/\tau_{sc}}}{\sum_{h'\neq h} e^{s(h,h')/\tau_{sc}}} \\ &= -\mathbb{E}_P\Big[s(h,h^+)/\tau_{sc}\Big] + \mathbb{E}\Big[\log\big(\sum_{h'\neq h} e^{s(h,h')/\tau_{sc}}\big)\Big]\end{aligned} \tag{1}$$

[1] https://stanfordnlp.github.io/CoreNLP

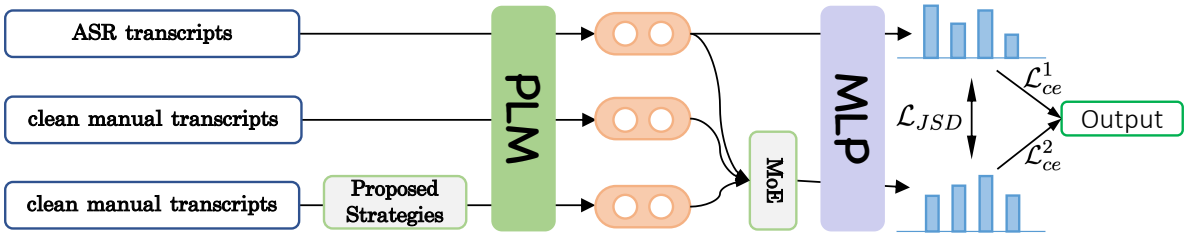

Figure 2: The illustration of the fine-tuning stage. We use the proposed strategies (§3.1) to generate more transcripts from clean manual transcripts and adopt mixture-of-experts (MoE) (§3.3) to weighted average the representations by pre-trained model (PLM) of generated transcripts, ASR transcripts, and the associated clean manual transcripts.

where P denotes the positive pairs, τ_{sc} denotes the temperature hyper-parameter to adjust the scale of the cosine similarity function $s(\cdot,\cdot)$. Positive pairs could be divided into the following three categories. The first category is ASR transcripts and associated clean manual transcripts, the second category is the ASR transcripts and the associated generated transcripts, and the third category is the clean manual transcripts and the associated generated transcripts. The negative pairs are all the remaining transcripts in the same batch, including ASR transcripts, clean manual transcripts, and the generated transcripts.

3.3 Mixture-of-Experts

As shown in Figure 2, to leverage the different characteristics of the clean manual transcripts and ASR transcripts more effectively, we design a mixture-of-experts module in fine-tuning. For a transcript $x^i = (x_1^i, x_2^i, \ldots, x_m^i)$, we concatenate these representations obtained by the pre-trained model and feed the concatenation to a multi-layer perceptron (MLP) to predict the confidence score β^i:

$$\beta^i = \tanh([h^i; h_1^i; \ldots; h_m^i]\mathbf{W}^i + b^i) \quad (2)$$

where $\mathbf{W}^i$ and b^i are the parameters of MLP, h^i denotes the representation of the special token `[CLS]`, $h_1^i, \ldots, h_m^i$ denote the representations of the original tokens, $[\cdot]$ denotes the concatenation operation.

For the input clean manual transcript x^1 and its associated ASR transcript x^2, we employ repetition, deletion, and swap strategies to obtain x^3, x^4, and x^5, respectively. We consider them as five experts and then apply a softmax function to calculate the weight G^i of the i-th expert:

$$G^i = \text{softmax}(\beta^i) \quad (3)$$

Then, we utilize x^1, x^2, x^3, x^4, and x^5 as inputs to the model and calculate the weighted average of their predictions, denoted as $\hat{y}$:

$$\hat{y} = \sum_{i=1}^{5} G^i \cdot \mathbf{F}(x^i) \quad (4)$$

where $\mathbf{F}(\cdot)$ denotes the corresponding predictions of the SLU model. Through utilizing this approach, the SLU model can assign different weights to the generated transcripts, manual transcripts, and ASR transcripts, effectively incorporating their respective strengths within the overall system.

3.4 Training Objective

Pre-training Motivated by recent success of pre-trained models (Cao et al., 2022; Li et al., 2022b, 2023; Jin et al., 2023; Dong et al., 2023b; Yang et al., 2024; Huang et al., 2024a), we proceed with training the masked language model (MLM) during the pre-training stage. Following Chang and Chen (2022), the training loss $\mathcal{L}_{pt}$ is the weighted sum of self-supervised contrastive learning loss $\mathcal{L}_{sc}$ and an MLM loss $\mathcal{L}_{mlm}$ as follows:

$$\mathcal{L}_{pt} = \lambda\mathcal{L}_{sc} + (1-\lambda)\cdot\mathcal{L}_{mlm} \quad (5)$$

where λ is the coefficient balancing the two tasks.

Fine-tuning Following E et al. (2019); Chen et al. (2022a), the intent detection objective is:

$$\mathcal{L}_I^1 = -\sum_{i=1}^{N} y_i^I \log y_i^{2,I} \quad (6)$$

$$\mathcal{L}_I^2 = -\sum_{i=1}^{N} y_i^I \log \hat{y}_i{}^I \quad (7)$$

where y_i^I denotes the intent label, $y_i^{2,I}$ is the prediction of the original ASR transcript, $\hat{y}_i{}^I$ denotes the weighted average of the intent predictions, N denotes the batch size, and n denotes the number of the tokens in the utterance.

For slot filling, we follow Dong et al. (2023c) to perform it in the sequence generation style, and the slot filling objective is as follows:

$$\mathcal{L}_S^1 = -\sum_{i=1}^{N}\sum_{j=1}^{n} y_i^{j,S} \log y_i^{2,j,S} \tag{8}$$

$$\mathcal{L}_S^2 = -\sum_{i=1}^{N}\sum_{j=1}^{n} y_i^{j,S} \log \hat{y_i}^{j,S} \tag{9}$$

where $y_i^{j,S}$ denotes the slot label, $y_i^{2,j,S}$ denotes the prediction of the original ASR transcript, and $\hat{y_i}^{j,S}$ denotes the weighted average of slot predictions.

We also regularize the weighted average of the predictions and the predictions of ASR transcripts by minimizing JSD between output distributions:

$$\mathcal{L}_{JSD}^I = \sum_{i=1}^{N} \mathrm{JSD}(y_i^{2,I}, \hat{y_i}^I) \tag{10}$$

$$\mathcal{L}_{JSD}^S = \sum_{i=1}^{N}\sum_{j=1}^{n} \mathrm{JSD}(y_i^{2,j,S}, \hat{y_i}^{j,S}) \tag{11}$$

The final fine-tuning loss $\mathcal{L}_{ft}$ is as follows:

$$\mathcal{L}_{ft} = \mathcal{L}_I^1 + \mathcal{L}_I^2 + \mathcal{L}_S^1 + \mathcal{L}_S^2 + \gamma(\mathcal{L}_{JSD}^I + \mathcal{L}_{JSD}^S) \tag{12}$$

where γ is the coefficient weight.

4 Experiments

4.1 Datasets and Metrics

Following previous works (Chang and Chen, 2022), all the experiments are conducted on three widely-used benchmark datasets, including SLURP, ATIS, and TREC6[2]. Table 1 presents the statistical information of the three datasets.

Dataset	#Class	Avg. Length	Train	Test
SLURP	18×46	6.93	50,628	10,992
ATIS	22	11.14	4,978	893
TREC6	6	8.89	5,452	500

Table 1: The statistics of all datasets. The *test* set of SLURP is sub-sampled.

SLURP is a challenging SLU dataset, which encompasses diverse domains, speakers, and recording settings. Each intent in SLURP is represented as a (scenario, action) pair, and the evaluation metric for SLURP is joint accuracy, which considers a prediction correct only if both the scenario and action are accurately predicted. The corresponding ASR transcripts are obtained using the Google Web API[3]. ATIS and TREC6 are two additional SLU datasets for flight reservation and question classification tasks, respectively. For these two datasets, the data synthesis involves a text-to-speech (TTS) model followed by ASR transcription.

4.2 Baselines

We compare our MoE-SLU with six pipeline SLU baselines, including RoBERTa (Liu et al., 2019), Phoneme-BERT (Sundararaman et al., 2021), SimCSE (Gao et al., 2021), and SpokenCSE (Chang and Chen, 2022), and MCLF (Huang et al., 2023), and ML-LMCL (Cheng et al., 2023a), and five end-to-end SLU baselines, including MTL-SLT (Huang et al., 2022), Speech-Brain (Ravanelli et al., 2021), CTI (Seo et al., 2022), HuBERT SLU (Wang et al., 2021), CIF-PT (Dong et al., 2023c). For a fair comparison, we report the results of MoE-SLU using RoBERTa (Liu et al., 2019) and Data2vec (Baevski et al., 2022) as the pre-trained models. The performance of recent large language models including ChatGPT[4](OpenAI, 2023) and SpeechGPT (Zhang et al., 2023) are also reported for comparison.

4.3 Implementation Details

During all the experiments, the model is pre-trained for 10k steps on each dataset utilizing a batch size 128. To avoid overfitting, the model is fine-tuned for up to 10 epochs with the batch size 256. Early stop is employed if the loss on *dev* set does not decrease for 3 epochs. For the SLURP dataset, two separate classification heads are trained for the scenario and the action, sharing the same embeddings. For all the hyper-parameters, we conduct several experiments and choose the values which perform best. The mask ratio of the MLM task is set to 0.15, τ_{sc} is set to 0.2, λ is set to 0.4, and γ is set to 2. During both pre-training and fine-tuning, we utilize the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98$, and 4k warm-up updates to optimize the parameters. The whole training process typically lasts a few hours. All the experiments are conducted on an Nvidia Tesla-A100 GPU.

[2]The SLURP dataset can be accessed at https://github.com/MiuLab/SpokenCSE, while the ATIS and TREC6 datasets can be accessed at https://github.com/Observe ai-Research/Phoneme-BERT.

[3]https://cloud.google.com/speech-to-text

[4]https://chat.openai.com

Model	Backbone	SLURP		ATIS		TREC6	
		Slot	Intent	Slot	Intent	Slot	Intent
Pipeline SLU Models							
RoBERTa (Liu et al., 2019)	RoBERTa	74.55♣	84.42	91.22♣	94.86	74.82♣	84.54
Phoneme-BERT (Sundararaman et al., 2021)	RoBERTa	74.52♣	84.16	92.24♣	95.14	76.33♣	86.48
SimCSE (Gao et al., 2021)	RoBERTa	74.59♣	84.88	91.87♣	94.32	75.44♣	85.46
SpokenCSE (Chang and Chen, 2022)	RoBERTa	74.63♣	85.64	92.81♣	95.58	76.97♣	86.82
MCLF (Huang et al., 2023)	RoBERTa	74.89♣	85.39	92.31♣	95.22	77.31♣	87.00
ML-LMCL (Cheng et al., 2023a)	RoBETRa	78.28♣	89.16	94.18♣	97.21	77.59♣	89.96
End-to-End SLU Models							
MTL-SLT (Huang et al., 2022)	LAS + BART	74.49	83.10	93.65	97.13	-	-
Speech-Brain (Ravanelli et al., 2021)	wav2vec 2.0	74.62	85.34	-	-	-	-
CTI (Seo et al., 2022)	wav2vec 2.0 + RoBERTa	74.66	86.92	-	-	-	-
HuBERT SLU (Wang et al., 2021)	HuBERT	78.92	89.38	-	-	-	-
CIF-PT (Dong et al., 2023c)	Conformer	78.67	89.60	-	-	-	-
CIF-PT (Dong et al., 2023c)	Data2vec	81.63	91.32	-	-	-	-
Large Language Models							
ChatGPT (OpenAI, 2023)		62.83	73.96	81.16	84.13	67.56	73.68
SpeechGPT (Zhang et al., 2023)		61.56	72.84	79.28	83.21	66.12	71.34
Ours							
MoE-SLU	RoBERTa	**79.25**†	**89.92**†	**94.59**†	**97.73**†	**79.23**†	**90.43**†
MoE-SLU	Data2vec	**82.71**†	**92.45**†	**94.92**†	**98.26**†	**79.67**†	**91.82**†

Table 2: Results of slot filling and intent detection on three datasets. '†' denotes MoE-SLU obtains statistically significant improvements over baselines with $p < 0.01$. '♣' and '-' indicate that the results are not available in the original papers and '♣' indicates that the results are obtained based on our implementation.

4.4 Main Results

The performance comparison between MoE-SLU and baselines is presented in Table 2. Based on the results, we have the following observations:

(1) When RoBERTa is selected as the pre-trained model, our MoE-SLU outperforms all the baselines except CIF-PT which applies Data2vec as the pre-trained model. Besides, when Data2vec is used as the pre-trained model, the performance of our MoE-SLU further boosts and surpasses all the baselines. This improvement can be attributed to the proposed strategies and the use of different weights assigned to generated transcripts, ASR transcripts, and clean manual transcripts. These distinct weights could enable the effective leveraging of the unique strengths of different types of transcript.

(2) We adopt the evaluation method introduced by He and Garner (2023) to assess the performance of ChatGPT and SpeechGPT using ASR transcripts, where the model is presented with 20 examples and prompted to accommodate for ASR errors. From the results, we can obviously observe that there is a performance gap of approximately 20% between these models and MoE-SLU on the SLURP dataset. This performance degradation can also be observed in other datasets. This discrepancy highlights the challenges that language models might encounter when it comes to comprehending spoken language in the presence of noise. Therefore, enhancing the robustness of LLMs to ASR input errors remains a highly valuable area of exploration.

4.5 Analysis

In this section, several analytical experiments are conducted. To maintain the fair comparison, we report the performance of MoE-SLU with RoBERTa as the pre-trained model unless stated otherwise.

4.5.1 Ablation Study

To provide evidence for the advantages of our MoE-SLU model from multiple perspectives, we conduct a set of ablation experiments on MoE-SLU. The experiment results are shown in Table 3.

Effectiveness of the Proposed Strategies. One of the core contributions of MoE-SLU is the proposed three strategies, which enlarges the training set. To evaluate the effectiveness of the three strategies, we conduct three ablation experiments where each strategy is individually removed. We denote the processed results without each strategy as "w/o

Model	SLURP		ATIS		TREC6	
	Slot	Intent	Slot	Intent	Slot	Intent
MoE-SLU	79.25	89.92	94.59	97.73	79.23	90.43
w/o Repetition Strategy	78.71 (↓0.54)	89.29 (↓0.63)	94.18 (↓0.41)	97.21 (↓0.52)	79.01 (↓0.22)	90.07 (↓0.36)
w/o Deletion Strategy	78.52 (↓0.73)	89.09 (↓0.83)	93.91 (↓0.68)	96.99 (↓0.74)	78.78 (↓0.45)	89.92 (↓0.51)
w/o Swap Strategy	78.84 (↓0.41)	89.37 (↓0.55)	94.27 (↓0.32)	97.27 (↓0.46)	79.07 (↓0.16)	90.15 (↓0.29)
w/o MoE	77.99 (↓1.26)	88.24 (↓1.68)	93.37 (↓1.22)	96.35 (↓1.38)	78.07 (↓1.16)	89.11 (↓1.32)
w/o MoE + batch size↑	78.03 (↓1.22)	88.26 (↓1.66)	93.43 (↓1.16)	96.41 (↓1.32)	78.11 (↓1.12)	89.16 (↓1.27)
w/o JSD	78.51 (↓0.74)	89.11 (↓0.81)	93.96 (↓0.63)	97.02 (↓0.71)	78.78 (↓0.45)	89.94 (↓0.49)

Table 3: Results of the ablation experiments when RoBERTa is selected as the pre-trained model.

Repetition Strategy", "w/o Deletion Strategy", and "w/o Swap Strategy" in Table 3, respectively. When any of the strategies is removed, we observe a significant decrease. These results verify that the proposed strategies can make a positive contribution to ASR-Robust SLU. The reason is that these strategies simulate more types of noise in the real-world environments and improve the ability of the model to capture the keywords, thereby indirectly enhancing the positive effects of mixture-of-experts.

Effectiveness of Mixture-of-Experts. Another core contribution of our MoE-SLU is the utilization of the mixture-of-experts approach, which is designed to effectively leverage the different transcripts. To validate the effectiveness of mixture-of-experts, we conduct an ablation experiment where we remove $\mathcal{L}_I^2$, $\mathcal{L}_S^2$, $\mathcal{L}_{JSD}^I$, and $\mathcal{L}_{JSD}^S$ in Eq.12, denoted as "w/o MoE" in Table 3. We could also find the obvious performance degradation. We believe the reason is that through learning the weights in mixture-of-experts, the model can benefit from the distinctive strengths of different transcripts.

Contrastive learning benefits from a larger batch size as it provides more negative examples to facilitate convergence (Chen et al., 2020). To delve into whether the proposed mixture-of-experts approach, rather than the indirectly increased batch sizes, is responsible for the improvements, we double the original batch size after not utilizing mixture-of-experts, denoted as "w/o MoE + batch size↑" in Table 3. The results show that these improvements are indeed attributed to the proposed mixture-of-experts approach rather than the boosted batch size.

Effectiveness of JSD. To verify the effectiveness of JSD, we also remove $\mathcal{L}_{JSD}^I$ and $\mathcal{L}_{JSD}^S$ in Eq.12 and denoted it as *w/o JSD* in Table 3. The results demonstrate that the accuracy drops by 0.81, 0.71, and 0.49 in intent and 0.74, 0.63, and 0.45 in slot on the three datasets, respectively, providing the evidence that JSD further enhances the performance. We suggest that the reason is that JSD serves as a regularization mechanism that further promotes the leveraging of the strengths inherent in different transcripts and helps to prevent overfitting during the entire training process.

4.5.2 Weight Analysis

We also report the learned weights of x^1, x^2, x^3, x^4, and x^5 in Eq. 4 on these three datasets. It is evident that the transcripts after the deletion strategy carry significant weight on all the datasets, aligning with the findings mentioned in Sec. 4.5.1. Furthermore, the transcripts after applying the repetition strategy and the swap strategy retain certain weights, providing further validation of the effectiveness of our proposed three data augmentation strategies.

Dataset	G^1	G^2	G^3	G^4	G^5
SLURP	0.13	0.18	0.21	0.31	0.17
ATIS	0.15	0.17	0.22	0.33	0.13
TREC6	0.14	0.17	0.23	0.34	0.12

Table 4: The learned weights in Eq. 4 on three datasets.

4.5.3 Visualization

To gain a deeper understanding of the impact and contributions of the mixture-of-experts method, we provide a visualization example on SLURP dataset in Figure 3 based on Principal Component Analysis (PCA) (Abdi and Williams, 2010). In our MoE-SLU approach, the representations of clean manual transcripts, transcripts after applying the proposed strategies, and their corresponding ASR transcripts are closely aligned, showcasing the robustness of our method in handling ASR errors. However, in the case of the previous best method, ML-LMCL,

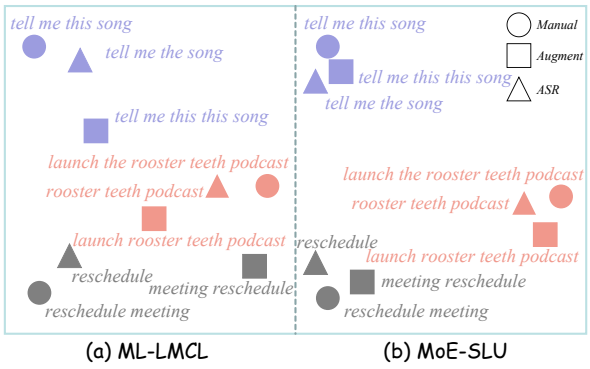

Figure 3: Visualization of representations of clean manual transcripts, transcripts after applying proposed strategies, and ASR transcripts. Circles, squares, and triangles in the same color indicate the corresponding transcriptions are associated.

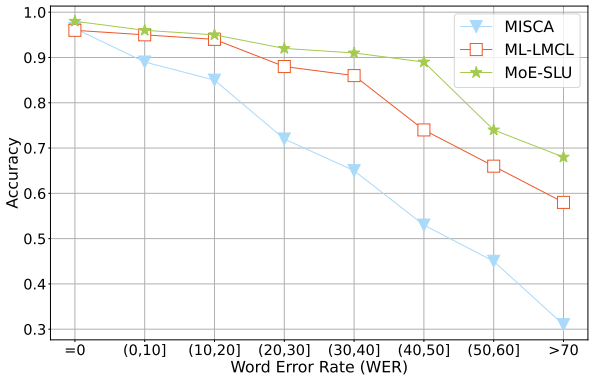

Figure 4: Accuracy of MISCA, ML-LMCL, and MoE-SLU at different levels of WER.

although the representations of clean manual transcripts and the corresponding ASR transcripts are also relatively close, they are distant from the associated transcripts after using the proposed strategies, which further supports that our MoE-SLU can effectively capture and utilize the unique information provided by different transcripts, resulting in the improved performance in aligning and comprehending the spoken language.

4.5.4 Performance at Different Noise Levels

To better investigate the impact of different noise levels, we separate the *test* set of SLURP dataset into eight groups based on their WER and demonstrate the intent accuracy in Figure 4. We choose a previous ASR-Robust SLU model ML-LMCL and an SLU model MISCA (Pham et al., 2023) which is not specially designed for improving ASR robustness. We can obviously observe that MISCA is significantly influenced by ASR errors, which proves the necessity of developing the ASR-Robust SLU frameworks. Compared to ML-LMCL, MoE-SLU indeed improves ASR Robustness more effectively, which further verifies the superiority of our method. We believe the reason is that our method leverages different strengths of each type of transcript.

5 Conclusion

In this paper, we propose a new ASR-Robust SLU framework MoE-SLU. We design three strategies to simulate more kinds of errors in the real environment and apply mixture-of-experts to leverage the different transcripts more effectively. Experiments and analysis on three datasets show that our model significantly outperforms previous models. Future work will focus on exploring how to make better use of the unique strengths of different transcripts.

Limitations

While MoE-SLU demonstrates substantial improvements over existing ASR-Robust SLU models, it is important to note that as previous works (Chang and Chen, 2022), we currently still rely on the ASR transcripts for pre-training and fine-tuning to align with the target inference scenario. However, ASR transcripts may not always be readily available due to the constraints of ASR systems. Therefore, our future research will aim to improve ASR robustness without relying on any ASR transcripts throughout the training and inference process.

Ethics Statement

We perform all experiments applying publicly available datasets that have been pre-processed for academic research purposes. As a result, these datasets do not contain any information which could identify individuals by name or contain offensive content. Though our framework could achieve state-of-the-art performance, it is essential to acknowledge that the results generated by our SLU framework might not be entirely perfect. So it is advisable for individuals not to depend solely on the generated results. In real-world applications, it is recommended to be assisted by humans to ensure accuracy.

Acknowledgements

This paper was partially supported by NSFC (No: 62176008). We thank all the anonymous reviewers for their insightful comments.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proc. of ICML*.

Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. In *Proc. of EMNLP*.

Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. 2022. Locvtp: Video-text pre-training for temporal localization. In *Proc. of ECCV*.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. of ICASSP*.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding. In *Proc. of Interspeech*.

Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022a. Towards joint intent detection and slot filling via higher-order attention. In *Proc. of IJCAI*.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2024. On the pareto front of multilingual neural machine translation. *Proc. of NeurIPS*.

Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022b. Joint multiple intent detection and slot filling via self-distillation. In *Proc. of ICASSP*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proc. of ICML*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Proc. of ACL Findings*.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023b. Mrrl: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Proc. of EMNLP Findings*.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, Xianwei Zhuang, and Yuexian Zou. 2024. Towards multi-intent spoken language understanding via hierarchical attention and optimal transport. In *Proc. of AAAI*.

Xuxin Cheng, Zhihong Zhu, Wanshi Xu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023c. Accelerating multiple intent detection and slot filling via targeted knowledge distillation. In *Proc. of EMNLP*.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. In *Proc. of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Luis Fernando D'Haro and Rafael E. Banchs. 2016. Automatic correction of ASR outputs by using machine translation. In *Proc. of INTERSPEECH*.

Guanting Dong, Daichi Guo, Liwen Wang, Xuefeng Li, Zechen Wang, Chen Zeng, Keqing He, Jinzheng Zhao, Hao Lei, Xinyue Cui, Yi Huang, Junlan Feng, and Weiran Xu. 2022. PSSAT: A perturbed semantic structure awareness transferring method for perturbation-robust slot filling. In *Proc. of COLING*.

Guanting Dong, Tingfeng Hui, Zhuoma GongQue, Jinxu Zhao, Daichi Guo, Gang Zhao, Keqing He, and Weiran Xu. 2023a. Demonsf: A multi-task demonstration-based generative framework for noisy slot filling task. In *Proc. of EMNLP Findings*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023b. How abilities in large language models are affected by supervised fine-tuning data composition. *ArXiv preprint*.

Linhao Dong, Zhecheng An, Peihao Wu, Jun Zhang, Lu Lu, and Ma Zejun. 2023c. CIF-PT: Bridging speech and text representations for spoken language understanding via continuous integrate-and-fire pre-training. In *Proc. of ACL Findings*.

Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Souvik Pal, Ganesh Ramakrishnan, and Preethi Jyothi. 2022. Error correction in asr using sequence-to-sequence models. *ArXiv preprint*.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proc. of ACL*.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. 2023. Language models can be logical solvers. *ArXiv preprint*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proc. of EMNLP*.

Patrick Haffner, Gokhan Tur, and Jerry H Wright. 2003. Optimizing svms for complex call classification. In *Proc. of ICASSP*.

Mutian He and Philip N. Garner. 2023. Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding. In *Proc. of Interspeech*.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2018. Sequence to sequence mixture model for diverse machine translation. In *Proc. of CoNLL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*.

Guanhua Huang, Zeping Min, Qian Ge, and Zhouwang Yang. 2024a. Towards document-level event extraction via binary contrastive generation. *Knowledge-Based Systems*.

Jianheng Huang, Ante Wang, Linfeng Gao, Linfeng Song, and Jinsong Su. 2024b. Response enhanced semi-supervised dialogue query generation. In *Proc. of AAAI*.

Zhiqi Huang, Dongsheng Chen, Zhihong Zhu, and Xuxin Cheng. 2023. Mclf: A multi-grained contrastive learning framework for asr-robust spoken language understanding. In *Proc. of EMNLP Findings*.

Zhiqi Huang, Milind Rao, Anirudh Raju, Zhe Zhang, Bach Bui, and Chul Lee. 2022. MTL-SLT: Multi-task learning for spoken language tasks. In *Proceedings of the 4th Workshop on NLP for Conversational AI*.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computing*.

Chuhao Jin, Yutao Zhu, Lingzhen Kong, Shijie Li, Xiao Zhang, Ruihua Song, Xu Chen, Huan Chen, Yuchong Sun, Yu Chen, et al. 2023. Joint semantic and strategy matching for persuasive dialogue. In *Proc. of EMNLP Findings*.

Michael I. Jordan and Robert A. Jacobs. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computing*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu. 2021. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. In *Proc. of NeurIPS*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *Proc. of SIGIR*.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022b. The past mistake is the future wisdom: Error-driven contrastive probability optimization for Chinese spell checking. In *Proc. of ACL Findings*.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023. Pace: Unified multi-modal dialogue pre-training with progressive and compositional experts. In *Proc. of ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. ASR error correction and domain adaptation using machine translation. In *Proc. of ICASSP*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL*.

OpenAI. 2023. Chatgpt.

Hao Peng, Roy Schwartz, Dianqi Li, and Noah A. Smith. 2020. A mixture of h - 1 heads is better than h heads. In *Proc. of ACL*.

Thinh Pham, Chi Tran, and Dat Quoc Nguyen. 2023. MISCA: A joint model for multiple intent detection and slot filling with intent-slot co-attention. In *Proc. of EMNLP Findings*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *ArXiv preprint*.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Proc. of Interspeech*.

Seunghyun Seo, Donghyun Kwak, and Bowon Lee. 2022. Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding. In *Proc. of ICASSP*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proc. of ICLR*.

Xiangqing Shen, Yurun Song, Siwei Wu, and Rui Xia. 2024. Vcd: Knowledge base guided visual commonsense discovery in images. *ArXiv preprint*.

Feifan Song, Lianzhe Huang, and Houfeng Wang. 2024. A unified framework for multi-intent spoken language understanding with prompting. In *Proc. of ICASSP*.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phonemebert: Joint language modelling of phoneme sequence and ASR transcript. In *Proc. of INTERSPEECH*.

Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proc. of ICML*.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *ArXiv preprint*.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. *ArXiv preprint*.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proc. of EMNLP*.

Yifei Xin, Xiulian Peng, and Yan Lu. 2023a. Masked audio modeling with clap and multi-objective learning. In *Proc. of Interspeech*.

Yifei Xin, Dongchao Yang, and Yuexian Zou. 2022. Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification. In *Proc. of Interspeech*.

Yifei Xin, Dongchao Yang, and Yuexian Zou. 2023b. Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss. In *Proc. of ICASSP*.

Yifei Xin and Yuexian Zou. 2023. Improving audio-text retrieval via hierarchical cross-modal interaction and auxiliary captions. In *Proc. of Interspeech*.

Puyang Xu and Ruhi Sarikaya. 2013a. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*.

Puyang Xu and Ruhi Sarikaya. 2013b. Exploiting shared information for multi-intent natural language sentence classification. In *Proc. of Interspeech*.

Weiyuan Xu, Peilin Zhou, Chenyu You, and Yuexian Zou. 2021. Semantic transportation prototypical network for few-shot intent detection. In *Proc. of INTERSPEECH*.

Yifan Yang, Jiajun Zhou, Ngai Wong, and Zheng Zhang. 2024. Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In *Proc. of AACL*.

Yongkang Yin, Xu Li, Ying Shan, and Yuexian Zou. 2023. Afl-net: Integrating audio, facial, and lip modalities with cross-attention for robust speaker diarization in the wild. *ArXiv preprint*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Proc. of EMNLP Findings*.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. In *Proc. of EMNLP*.

Jinxu Zhao, Guanting Dong, Yueyan Qiu, Tingfeng Hui, Xiaoshuai Song, Daichi Guo, and Weiran Xu. 2024. Noise-bert: A unified perturbation-robust framework with noise alignment pre-training for noisy slot filling task. In *Proc. of ICASSP*.

Peilin Zhou, Dading Chong, Helin Wang, and Qingcheng Zeng. 2022. Calibrate and refine! a novel and agile framework for asr-error robust intent detection. In *Proc. of Interspeech*.

Peilin Zhou, Zhiqi Huang, Fenglin Liu, and Yuexian Zou. 2021. Pin: A novel parallel interactive network for spoken language understanding. In *Proc. of ICPR*.

Xiner Zhu, Yichao Wu, Haoji Hu, Xianwei Zhuang, Jincao Yao, Di Ou, Wei Li, Mei Song, Na Feng, and Dong Xu. 2022. Medical lesion segmentation by combining multimodal images with modality weighted unet. *Medical Physics*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023a. Enhancing code-switching for cross-lingual slu: a unified view of semantic and grammatical coherence. In *Proc. of EMNLP*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023b. Towards unified spoken language understanding decoding via label-aware compact linguistics representations. In *Proc. of ACL Findings*.

Zhihong Zhu, Xuxin Cheng, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2024. Aligner^2: Enhancing joint multiple intent detection and slot filling via adjustive and forced cross-task alignment. In *Proc. of AAAI*.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. In *Proc. of AAAI*.

Xianwei Zhuang, Xiner Zhu, Haoji Hu, Jincao Yao, Wei Li, Chen Yang, Liping Wang, Na Feng, and Dong Xu. 2022. Residual swin transformer unet with consistency regularization for automatic breast ultrasound tumor segmentation. In *Proc. of ICIP*.