ML-LMCL: Mutual Learning and Large-Margin Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding

Xuxin Cheng, Bowen Cao[†], Qichen Ye[†],
Zhihong Zhu[†], Hongxiang Li, Yuexian Zou[*]
School of ECE, Peking University, China
{chengxx, cbw2021, zhihongzhu, lihongxiang}@stu.pku.edu.cn
{yeeeqichen, zouyx}@pku.edu.cn

Abstract

Spoken language understanding (SLU) is a fundamental task in the task-oriented dialogue systems. However, the inevitable errors from automatic speech recognition (ASR) usually impair the understanding performance and lead to error propagation. Although there are some attempts to address this problem through contrastive learning, they (1) treat clean manual transcripts and ASR transcripts equally without discrimination in fine-tuning; (2) neglect the fact that the semantically similar pairs are still pushed away when applying contrastive learning; (3) suffer from the problem of Kullback–Leibler (KL) vanishing. In this paper, we propose **M**utual **L**earning and **L**arge-**M**argin **C**ontrastive **L**earning (ML-LMCL), a novel framework for improving ASR robustness in SLU. Specifically, in fine-tuning, we apply mutual learning and train two SLU models on the manual transcripts and the ASR transcripts, respectively, aiming to iteratively share knowledge between these two models. We also introduce a distance polarization regularizer to avoid pushing away the intra-cluster pairs as much as possible. Moreover, we use a cyclical annealing schedule to mitigate KL vanishing issue. Experiments on three datasets show that ML-LMCL outperforms existing models and achieves new state-of-the-art performance.

1 Introduction

Spoken language understanding(SLU) is an important component of various personal assistants, such as Amazon's Alexa, Apple's Siri, Microsoft's Cortana and Google's Assistant (Young et al., 2013). SLU aims at taking human speech input and extracting semantic information for two typical subtasks, mainly including intent detection and slot filling (Tur and De Mori, 2011). Pipeline approaches and end-to-end approaches are two kinds of solutions of SLU. Pipeline SLU methods usually combine automatic speech recognitgion (ASR) and natural language understanding (NLU) in a cascaded manner, so they can easily apply external datasets and external pre-trained language models. However, error propagation is a common problem of pipeline approaches, where an inaccurate ASR output can theoretically lead to a series of errors in subtasks. As shown in Figure 1, due to the error from ASR, the model can not predict the intent correctly. Following Chang and Chen (2022), this paper only focuses on intent detection.

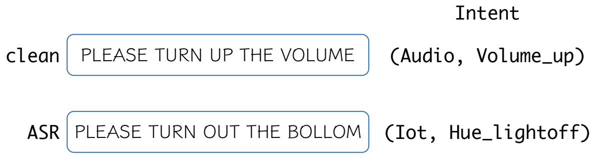


Figure 1: An example of the intent being predicted incorrectly due to the ASR error.

Learning error-robust representations is an effective method to mitigate the negative impact of errors from ASR and is gaining increasing attention. The remedies for ASR errors can be broadly categorized into two types: (1) applying machine translation to translate the erroneous ASR transcripts to clean manual transcripts (Mani et al., 2020; Wang et al., 2020; Dutta et al., 2022); (2) using masked language modeling to adapt the model. However, these methods usually requires additional speech-related inputs (Huang and Chen, 2019; Sergio et al., 2020; Wang et al., 2022), which may not always be readily available. Therefore, this paper focuses on improving ASR robustness in SLU without using any speech-related input features.

Despite existing error-robust SLU models have achieved promising progress, we discover that they suffer from three main issues:

(1) **Manual and ASR transcripts are treated as the same type.** In fine-tuning, existing methods simply combine manual and ASR transcripts as the

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

Findings of the Association for Computational Linguistics: ACL 2023, pages 6492–6505
July 9-14, 2023 ©2023 Association for Computational Linguistics

final dataset, which limits the performance. Intuitively, the information from manual transcripts and the information from ASR transcripts play different roles, so the model fine-tuned on their combination cannot discriminate their specific contributions. Based on our observations, models trained on the clean manual transcripts usually has higher accuracy, while models trained on the ASR transcripts are usually more robust to ASR errors. Therefore, manual and ASR transcripts should be treated differently to improve the performance of the model.

(2) **Semantically similar pairs are still pushed away.** Conventional contrastive learning enlarges distances between all pairs of instances and potentially leading to some ambiguous intra-cluster and inter-cluster distances (Mishchuk et al., 2017; Zhang et al., 2022), which is detrimental for SLU. Specifically, if clean manual transcripts are pushed away from their associated ASR transcripts while become closer to other sentences, the negative impact of ASR errors will be further exacerbated.

(3) **They suffer from the problem of KL vanishing.** Inevitable label noise usually has a negative impact on the model (Li et al., 2022; Cheng et al., 2023b). Existing methods apply self-distillation to minimize Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) between the current prediction and the previous one to reduce the label noises in the training set. However, we find these methods suffer from the KL vanishing issue, which has been observed in other tasks (Zhao et al., 2017). KL vanishing can adversely affect the training of the model. Therefore, it is crucial to solve this problem to improve the performance.

In this paper, we propose **M**utual **L**earning and **L**arge-**M**argin **C**ontrastive **L**earning (ML-LMCL), a novel framework to tackle above three issues. For the first issue, we propose a mutual learning paradigm. In fine-tuning, we train two SLU models on the manual and ASR transcripts, respectively. These two models are collaboratively trained and considered as peers, with the aim of iteratively learning and sharing the knowledge between the two models. Mutual learning allows effective dual knowledge transfer (Liao et al., 2020; Zhao et al., 2021; Zhu et al., 2021), which can improve the performance. For the second issue, our framework implements a large-margin contrastive learning to distinguish between intra-cluster and inter-cluster pairs. Specifically, we apply a distance polarization regularizer and penalize all pairwise distances within the margin region, which can encourage polarized distances for similarity determination and obtain a large margin in the distance space in an unsupervised way. For the third issue, following Fu et al. (2019), we mitigate KL vanishing by adopting a cyclical annealing schedule. The training process is effectively split into many cycles. In each cycle, the coefficient of KL Divergence progressively increases from 0 to 1 during some iterations and then stays at 1 for the remaining iterations. Experiment results on three datasets SLURP, ATIS and TREC6 (Bastianelli et al., 2020; Hemphill et al., 1990; Li and Roth, 2002; Chang and Chen, 2022) demonstrate that our ML-LMCL significantly outperforms previous best models and model analysis further verifies the advantages of our model.

The contributions of our work are four-fold:

- We propose ML-LMCL, which utilizes mutual learning to encourage the exchange of knowledge between the model trained on clean manual transcripts and the model trained on ASR transcripts. To the best of our knowledge, we make the first attempt to apply mutual learning to improve ASR robustness in SLU task.
- To better distinguish between intra-cluster and inter-cluster pairs, we introduce a distance polarization regularizer to achieve large-margin contrastive learning.
- We adopt a cyclical annealing schedule to mitigate KL vanishing, which is neglected in the previous SLU approaches.
- Experiments on three public datasets demonstrate that the proposed model achieves new state-of-the-art performance.

2 Approach

Our framework includes four elements: (1) Self-supervised contrastive learning with a distance polarization regularizer in pre-training. (2) Mutual learning between the model trained on clean manual transcripts and the model trained on ASR transcripts in fine-tuning. (3) Supervised contrastive learning with a distance polarization regularizer in fine-tuning. (4) Self-distillation with the cyclical annealing schedule in fine-tuning.

2.1 Self-supervised Contrastive Learning

Following Chang and Chen (2022), we utilize self-supervised contrastive learning in pre-training for learning sentence representations invariant to misrecognition to handle ASR errors. Inspired by the

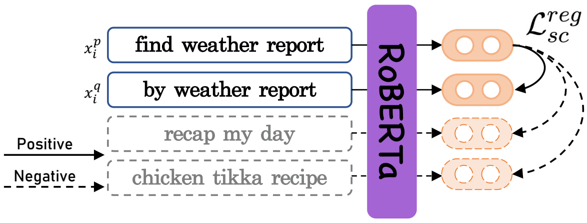


Figure 2: The illustration of the pre-training stage. We apply large-margin self-supervised contrastive learning with paired transcripts. A positive pair consists of clean data and the associated ASR transcript.

success of pre-trained models (Liu et al., 2022b; Zhang et al., 2023a; Cheng et al., 2023a; Zhang et al., 2023b; Yang et al., 2023a), we continually train a pre-trained RoBERTa (Liu et al., 2019) on spoken language corpus.

Given a mini-batch of input data of N pairs of transcripts $B = \{(x_i^p, x_i^q)\}_{i=1..N}$, where x_i^p denotes a clean manual transcript and x_i^q denotes its associated ASR transcript. As shown in Figure 2, we first apply the pre-trained RoBERTa and utilize the last layer of `[CLS]` to obtain the representation h_i^p for x_i^p and h_i^q for x_i^q:

$$h_i^p = \text{RoBERTa}(x_i^p) \tag{1}$$

$$h_i^q = \text{RoBERTa}(x_i^q) \tag{2}$$

Then we apply the proposed self-supervised contrastive loss $\mathcal{L}_{sc}$ (Chen et al., 2020a; Gao et al., 2021) to adjust the sentence representations:

$$\begin{aligned}\mathcal{L}_{sc} &= -\frac{1}{2N}\sum_{(h,h^+)\in P}\log\frac{e^{s(h,h^+)/\tau_{sc}}}{\sum_{h'\neq h}^{B} e^{s(h,h')/\tau_{sc}}}\\ &= -\mathbb{E}_P\Big[s(h,h^+)/\tau_{sc}\Big] + \mathbb{E}\Big[\log\big(\sum_{h'\neq h}^{B} e^{s(h,h')/\tau_{sc}}\big)\Big]\end{aligned} \tag{3}$$

where P is composed of $2N$ positive pairs of either (h_i^p, h_i^q) or (h_i^q, h_i^p), τ_{sc} is the temperature hyper-parameter and $s(\cdot,\cdot)$ denotes the cosine similarity function. In Eq.3, the first term brings the clean manual transcript and its associated ASR transcript (positive example) near together and the second term pushes irrelevant ones (negative examples) far apart to promote uniformity in representation space (Wang and Isola, 2020). Note that for a transcript, its negative examples may be clean manual transcripts or ASR transcripts. For example, in Figure 2, *recap my day* is a clean manual transcript and *chicken tikka recipe* is an ASR transcript.

However, conventional contrastive learning has a problem that semantically similar pairs are still pushed away (Chen et al., 2021). It indiscriminately enlarges distances between all pairs of instances and may not be able to distinguish intra-cluster and inter-cluster correctly, which causes some similar instance pairs to still be pushed away. Moreover, it may discard some negative pairs and regard them as semantically similar pairs wrongly, even though their learning objective treat each pair of original instances as dissimilar. These problems result in the distance between the clean manual transcript and its associated ASR transcript not being significantly smaller than the distance between unpaired instance, which is detrimental to improving ASR robustness. Motivated by Chen et al. (2021), we introduce a distance polarization regularizer to build a large-margin contrastive learning model. For simplicity, we further denote the following normalized cosine similarity:

$$\mathcal{D}_{ij} = (1 + s(h_i, h_j))\,/2 \tag{4}$$

which measures the similarity between the pairs of $(h_i, h_j) \in B$ with the real value $\mathcal{D}_{ij} \in [0, 1]$. We suppose that the matrix $\boldsymbol{\mathcal{D}} = \left\{\mathcal{D}_{ij} \in \mathbb{R}^{M\times M}\right\}$ where $M = 2N$ denotes the total number of transcripts in B. $\boldsymbol{\mathcal{D}}$ consists of distances $\mathcal{D}_{ij}$ and there exists $0 < \delta^+ < \delta^- < 1$ where the intra-class distances are smaller than δ^+ while the inter-class distances are larger than δ^-. The proposed distance polarization regularizer $\mathcal{L}_{reg}$ is as follows:

$$\mathcal{L}_{reg} = \left\|\min\left((\boldsymbol{\mathcal{D}} - \boldsymbol{\Delta}^+) \odot (\boldsymbol{\mathcal{D}} - \boldsymbol{\Delta}^-), 0\right)\right\|_1 \tag{5}$$

where $\boldsymbol{\Delta}^+ = \delta^+ \times \mathbf{1}_{M\times M}$ and $\boldsymbol{\Delta}^- = \delta^- \times \mathbf{1}_{M\times M}$ are the threshold parameters and $\|\cdot\|_1$ denotes the ℓ_1-norm. The region $(\delta^+, \delta^-) \subseteq [0, 1]$ can be regarded as the large margin to discriminate the similarity of data pairs. $\mathcal{L}_{reg}$ can encourage the sparse distance distribution in the margin region (δ^+, δ^-), because any distance $\mathcal{D}_{ij}$ fallen into the margin region (δ^+, δ^-) will increase $\mathcal{L}_{reg}$. Minimizing the regularizer $\mathcal{L}_{reg}$ will encourage more pairwise distances $\{\mathcal{D}_{ij}\}_{i,j=1}^{M}$ to distribute in the regions $[0, \delta^+]$ or $[\delta^-, 1]$, and each data pair is adaptively separated into similar or dissimilar result. As a result, through introducing the regularizer, our framework can better distinguish between intra-cluster and inter-cluster pairs.

Then the final large-margin self-supervised contrastive learning loss $\mathcal{L}_{sc}^{reg}$ is the weighted sum of self-supervised contrastive learning loss $\mathcal{L}_{sc}$ and the regularizer $\mathcal{L}_{reg}$, which is calculated as follows:

$$\mathcal{L}_{sc}^{reg} = \mathcal{L}_{sc} + \lambda_{reg}\cdot\mathcal{L}_{reg} \tag{6}$$

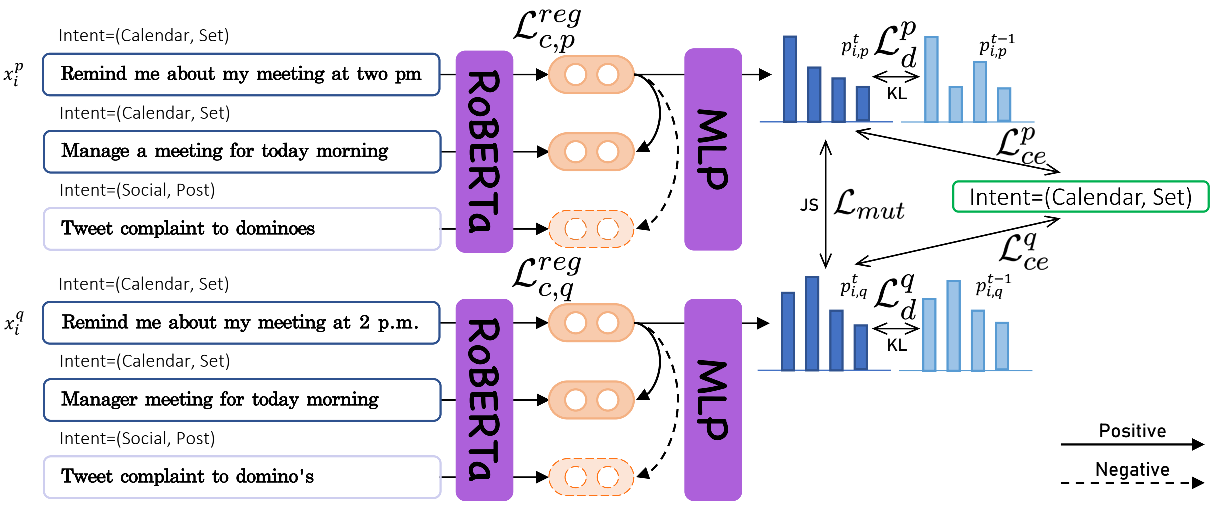

$\mathcal{L}_{c,p}^{reg}$
Intent=(Calendar, Set)
x_i^p
Remind me about my meeting at two pm
Intent=(Calendar, Set)
Manage a meeting for today morning
Intent=(Social, Post)
Tweet complaint to dominoes
RoBERTa
MLP
$p_{i,p}^t$ $\mathcal{L}_d^p$ $p_{i,p}^{t-1}$
KL
$\mathcal{L}_{ce}^p$
JS $\mathcal{L}_{mut}$
Intent=(Calendar, Set)
$\mathcal{L}_{ce}^q$
$\mathcal{L}_{c,q}^{reg}$
Intent=(Calendar, Set)
x_i^q
Remind me about my meeting at 2 p.m.
Intent=(Calendar, Set)
Manager meeting for today morning
Intent=(Social, Post)
Tweet complaint to domino's
RoBERTa
MLP
$p_{i,q}^t$ $\mathcal{L}_d^q$ $p_{i,q}^{t-1}$
KL
Positive
Negative

Figure 3: The illustration of the fine-tuning stage. Two networks on the clean manual transcripts and the ASR transcripts are collaboratively trained via mutual learning (§2.2). Large-margin supervised contrastive learning (§2.3) and self-distillation (§2.4) are applied to further reduce the impact of ASR errors.

where λ_{reg} is a hyper-parameter.

2.2 Mutual Learning

Previous work reveals that mutual learning can exploit the mutual guidance information between two models to improve their performance simultaneously (Nie et al., 2018; Hong et al., 2021). By mutual learning, we can obtain compact networks that perform better than those distilled from a strong but static teacher. In fine-tuning, we use the same pre-trained model in Sec.2.1 to train two networks on the manual transcripts and the ASR transcripts, respectively. For a manual transcript x_i^p and its associated ASR transcript x_i^q, the output probabilities $p_{i,p}^t$ and $p_{i,q}^t$ at the t-th epoch are as follows:

$$p_{i,p}^t = M_{\text{clean}}(x_i^p) \tag{7}$$

$$p_{i,q}^t = M_{\text{asr}}(x_i^q) \tag{8}$$

where M_{clean} denotes the model trained on clean manual transcripts and M_{asr} denotes the model trained on ASR transcripts.

We adopt Jensen-Shannon (JS) divergence as the mimicry loss, with the aim of effectively encouraging the two models to mimic each other. The mutual learning loss $\mathcal{L}_{mut}$ in Figure 3 is as follows:

$$\mathcal{L}_{mut} = \sum_{i=1}^{N} JS(p_{i,p}^t \| p_{i,q}^t) \tag{9}$$

2.3 Supervised Contrastive Learning

We also apply supervised contrastive learning in fine-tuning by using label information. The pairs with the same label are regarded as positive samples and the pairs with different labels are regarded as negative samples. The embeddings of positive samples are pulled closer while the embeddings of negative samples are pushed away (Jian et al., 2022; Zhou et al., 2022). We utilize the supervised contrastive loss $\mathcal{L}_c^p$ for the model trained on manual transcripts and $\mathcal{L}_c^q$ for the model trained on ASR transcripts to encourage the learned representations to be aligned with their labels:

$$\mathcal{L}_c^p = -\frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{j \neq i}^{N} 1_{y_i^p = y_j^p} \log \frac{e^{s(h_i^p, h_j^p)/\tau_c}}{\sum_{k \neq i}^{N} e^{s(h_i^p, h_k^p)/\tau_c}} \tag{10}$$

$$\mathcal{L}_c^q = -\frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{j \neq i}^{N} 1_{y_i^q = y_j^q} \log \frac{e^{s(h_i^q, h_j^q)/\tau_c}}{\sum_{k \neq i}^{N} e^{s(h_i^q, h_k^q)/\tau_c}} \tag{11}$$

where $y_i^p = y_j^p$ denotes the labels of h_i^p and h_j^p are the same, $y_i^q = y_j^q$ denotes the label of h_i^q and h_j^q are the same and τ_c is the temperature hyper-parameter.

Like Sec.2.1, we also use distance polarization regularizers $\mathcal{L}_{reg}^p$ and $\mathcal{L}_{reg}^q$ to enhance the generalization ability of contrastive learning algorithm:

$$\mathcal{L}_{reg}^p = \left\| \min\left(\left(\boldsymbol{\mathcal{D}}^p - \boldsymbol{\Delta}^+ \right) \odot \left(\boldsymbol{\mathcal{D}}^p - \boldsymbol{\Delta}^- \right), 0 \right) \right\|_1 \tag{12}$$

$$\mathcal{L}_{reg}^q = \left\| \min\left(\left(\boldsymbol{\mathcal{D}}^q - \boldsymbol{\Delta}^+ \right) \odot \left(\boldsymbol{\mathcal{D}}^q - \boldsymbol{\Delta}^- \right), 0 \right) \right\|_1 \tag{13}$$

where $\boldsymbol{\mathcal{D}}^p$ denotes the matrix consisting of pairwise distances on the clean manual transcripts and $\boldsymbol{\mathcal{D}}^q$ denotes the matrix on the ASR transcripts.

The large-margin supervised contrastive learning loss $\mathcal{L}_{c,p}^{reg}$ and $\mathcal{L}_{c,q}^{reg}$ in Figure 3 are as follows:

$$\mathcal{L}_{c,p}^{reg} = \mathcal{L}_c^p + \lambda_{reg}^p \mathcal{L}_{reg}^p \tag{14}$$

$$\mathcal{L}_{c,q}^{reg} = \mathcal{L}_c^q + \lambda_{reg}^q \mathcal{L}_{reg}^q \tag{15}$$

where λ_{reg}^{p} and λ_{reg}^{q} are two hyper-parameters.

The final large-margin supervised contrastive learning loss $\mathcal{L}_{c}^{reg}$ is as follows:

$$\mathcal{L}_{c}^{reg} = \mathcal{L}_{c,p}^{reg} + \mathcal{L}_{c,q}^{reg} \tag{16}$$

2.4 Self-distillation

To further reduce the impact of ASR errors, we apply a self-distillation method. We try to regularize the model by minimizing Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951; He et al., 2022) between the current prediction and the previous one (Liu et al., 2020, 2021). For the manual transcript x_i^p and its corresponding label y_i^p, $p_{i,p}^{t} = P(y_i^p|x_i^p, t)$ denotes the probability distribution of x_i^p at the t-th epoch, and $p_{i,q}^{t} = P(y_i^q|x_i^q, t)$ denotes the probability distribution of x_i^q at the t-th epoch. The loss functions $\mathcal{L}_d^p$ and $\mathcal{L}_d^q$ of self-distillation in Figure 3 are formulated as:

$$\mathcal{L}_d^p = \frac{1}{N}\sum_{i=1}^{N} \tau_d^2 KL\Big(\frac{p_{i,p}^{t-1}}{\tau_d} \| \frac{p_{i,p}^{t}}{\tau_d}\Big) \tag{17}$$

$$\mathcal{L}_d^q = \frac{1}{N}\sum_{i=1}^{N} \tau_d^2 KL\Big(\frac{p_{i,q}^{t-1}}{\tau_d} \| \frac{p_{i,q}^{t}}{\tau_d}\Big) \tag{18}$$

where τ_d is the temperature to scale the smoothness of two distributions, note that $p_{i,p}^{0}$ is the one-hot vector of label y_i^p and $p_{i,q}^{0}$ is that of label y_i^q.

Then the final self-distillation loss $\mathcal{L}_d$ is the sum of two loss functions $\mathcal{L}_d^p$ and $\mathcal{L}_d^q$:

$$\mathcal{L}_d = \mathcal{L}_d^p + \mathcal{L}_d^q \tag{19}$$

2.5 Training Objective

Pre-training Following (Chang and Chen, 2022), the pre-training loss $\mathcal{L}_{pt}$ is the weighted sum of the large-margin self-supervised contrastive learning loss $\mathcal{L}_{sc}^{reg}$ and an MLM loss $\mathcal{L}_{mlm}$:

$$\mathcal{L}_{pt} = \lambda_{pt}\mathcal{L}_{sc}^{reg} + (1-\lambda_{pt}) \cdot \mathcal{L}_{mlm} \tag{20}$$

where λ_{pt} is the coefficient balancing the two tasks.

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

$$\mathcal{L}_{ce}^{p} = -\sum_{i=1}^{N} y_i^p \log p_{i,p}^{t} \tag{21}$$

$$\mathcal{L}_{ce}^{q} = -\sum_{i=1}^{N} y_i^q \log p_{i,q}^{t} \tag{22}$$

$$\mathcal{L}_{ce} = \mathcal{L}_{ce}^{p} + \mathcal{L}_{ce}^{q} \tag{23}$$

The final fine-tuning loss $\mathcal{L}_{ft}$ is the weighted sum of cross-entropy loss L_{ce}, mutual learning loss $\mathcal{L}_{mut}$, large-margin supervised contrastive learning loss $\mathcal{L}_{c}^{reg}$ and self-distillation loss $\mathcal{L}_d$:

$$\mathcal{L}_{ft} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{mut} + \beta\mathcal{L}_{c}^{reg} + \gamma\mathcal{L}_d \tag{24}$$

where α, β, γ are the trade-off hyper-parameters.

However, directly using KL divergence for self-ditillation loss may suffer from the vanishing issue. To mitigate KL vanishing issue, we adopt a cyclical annealing schedule, which is also applied for this purpose in Fu et al. (2019); Zhao et al. (2021). Concretely, γ in Eq.24 changes periodically during training iterations, which is described by Eq.25:

$$\gamma = \begin{cases} \frac{r}{RC}, & r \leqslant RG \\ 1, & r > RG \end{cases} \tag{25}$$

$$r = mod(t-1, G) \tag{26}$$

where t represents the current training iteration and R and G are two hyper-parameters.

3 Experiments

3.1 Datasets and Metrics

Following Chang and Chen (2022), we conduct the experiments on three publicly available benchmark datasets[1]: SLURP, ATIS and TREC6 (Bastianelli et al., 2020; Hemphill et al., 1990; Li and Roth, 2002; Chang and Chen, 2022). The statistics of the three datasets included are shown in Table 1.

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 with various domains, speakers, and recording settings. An intent of SLURP is a (scenario, action) pair, the joint accuracy is used as the evaluation metric and the prediction is considered correct only when both scenario and action are correctly predicted. The ASR transcripts are obtained by Google Web API.

ATIS and TREC6 are two SLU datasets for flight reservation and question classification respectively.

[1] SLURP is available at https://github.com/MiuLab/SpokenCSE, and ATIS and TREC6 are available at https://github.com/Observeai-Research/Phoneme-BERT.

Model	w/o manual transcripts			w/ manual transcripts		
	SLURP	ATIS	TREC6	SLURP	ATIS	TREC6
RoBERTa (Liu et al., 2019)	83.97	94.53	84.08	84.42	94.86	84.54
Phoneme-BERT (Sundararaman et al., 2021)	83.78	94.83	85.96	84.16	95.14	86.48
SimCSE (Gao et al., 2021)	84.47	94.07	84.92	84.88	94.32	85.46
SpokenCSE (Chang and Chen, 2022)	85.26	95.10	86.36	85.64	95.58	86.82
ML-LMCL	**88.52**†	**96.52**†	**89.24**†	**89.16**†	**97.21**†	**89.96**†

Table 2: Accuracy results on three datasets. † denotes ML-LMCL obtains statistically significant improvements over baselines with p < 0.01. "w/o manual transcripts" denotes clean manual transcripts are not used in fine-tuning, *i.e.* the loss functions associated with clean manual transcripts are set to 0, including $\mathcal{L}_{ce}^{p}$, $\mathcal{L}_{mut}$, $\mathcal{L}_{c,p}^{reg}$, and $\mathcal{L}_{d}^{p}$. "w/ manual transcripts" denotes clean manual transcripts are used in fine-tuning.

We use the synthesized text released by Phoneme-BERT (Sundararaman et al., 2021), where the data is synthesized by a text-to-speech (TTS) model and later transcribed by ASR. We adopt accuracy as the evaluation metric for intent detection.

3.2 Implementation Details

We pre-train the model for 10K steps with a batch size 128 on each dataset, and finetune the whole model up to 10 epochs with a batch size 256 to avoid overfitting. The training will early-stop if the loss on *dev* set does not decrease for 3 epochs. On SLURP, two separate classification heads are trained for scenario and action with the shared BERT embeddings. The mask ratio of MLM is set to 0.15, τ_{sc} is set to 0.2, δ^+ is set to 0.2, δ^- is set to 0.5, λ_{reg} is set to 0.1, τ_c is set to 0.2, λ_{reg}^{p} is set to 0.15, λ_{reg}^{q} is set to 0.15, τ_d is set to 5, λ_{pt} is set to 0.5, α is set to 1, β is set to 0.1, R is set to 0.5, and G is set to 5000. The reported scores are averaged over 5 runs. During both pre-training and fine-tuning, we utilize 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 training process lasts a few hours. All experiments are conducted at an Nvidia Tesla-A100 GPU.

3.3 Baslines

We compare our model with the following baselines: (1) RoBERTa (Liu et al., 2019): a RoBERTa-base model directly fine-tuned on the target training data; (2) Phoneme-BERT (Sundararaman et al., 2021): a RoBERTa-base model which is further pre-trained on an additional corpus with the phoneme information and then fine-tuned on the target training data; (3) SimCSE (Gao et al., 2021): a state-of-the-art sentence embedding method applying contrastive learning; (4) SpokenCSE (Chang and Chen, 2022): a strong baseline for improving ASR robustness in SLU task.

3.4 Main Results

The performance comparison of ML-LMCL Net and baselines are shown in Table 2, from which we have the following observations:

(1) Our ML-LMCL gains consistent improvements on all tasks and datasets. This is because our model achieves the mutual guidance between the model trained on the manual and ASR transcripts, allowing these two models to share the knowledge for each other. Moreover, large-margin contrastive learning encourages the model to more accurately distinguish between intra-cluster and inter-cluster pairs, which can avoid pushing away the semantically similar pairs as much as possible. And cyclical annealing schedule is applied to mitigate KL vanish, which can improve the robustness of the model. When not using manual transcripts, it still overpasses SpokenCSE, which also demonstrates the effectiveness of large-margin contrastive learning and cyclical annealing schedule to improve ASR robustness in SLU.

(2) In contrast, it is obvious that the improvement on SLURP dataset is more significant. We believe the reason is that SLURP is a more challenging SLU dataset than ATIS and TREC6. An intent of SLURP is a (scenario, action) pair and the prediction is considered to be correct only if the scenario and action are both correctly predicted. Due to the shortcomings of conventional contrastive learning, previous work fail to align the ASR transcript and its associate manual transcript with high accuracy. As a result, due to ASR errors, it is common that one of the two components of an intent is incor-

rectly predicted. Our ML-LMCL is dedicated to overcome the shortcomings of conventional contrastive learning, resulting in better alignment and the improvement of performance.

3.5 Analysis

To verify the advantages of ML-LMCL from different perspectives, we use clean manual transcripts and conduct a set of ablation experiments. The experimental results are shown in Table 3.

Model	w/ manual transcripts		
	SLURP	ATIS	TREC6
ML-LMCL	**89.16**	**97.21**	**89.96**
w/o $\mathcal{L}_{mut}$	88.68 (↓0.48)	96.83 (↓0.38)	89.52 (↓0.44)
w/o $\mathcal{L}_{reg}$	88.92 (↓0.24)	96.98 (↓0.23)	89.77 (↓0.19)
w/o $\mathcal{L}^{q}_{reg}$ & $\mathcal{L}^{q}_{reg}$	88.75 (↓0.41)	96.92 (↓0.29)	89.74 (↓0.22)
w/o cyc	88.98 (↓0.18)	97.08 (↓0.13)	89.85 (↓0.11)
w/o L_{mut} + bsz↑	88.72 (↓0.44)	96.92 (↓0.29)	89.65 (↓0.31)
w/ $\mathcal{L}_{soft}$	89.12 (↓0.04)	97.18 (↓0.03)	89.92 (↓0.04)

Table 3: Results of the ablation experiments when using clean manual transcripts.

3.5.1 Effectiveness of Mutual Learning

One of the core contributions of ML-LMCL is mutual learning, which allows the two models trained on manual and ASR transcripts learn from each other. To verify the effectiveness of mutual learning, we remove mutual learning loss and refer it to *w/o* $\mathcal{L}_{mut}$ in Table 3. We observe that accuracy drops by 0.48, 0.38 and 0.44 on SLURP, ATIS and TREC6, respectively. Contrastive learning benefits more from larger batch size because larger batch size provides more negative examples to facilitate convergence (Chen et al., 2020a), and many attempts have been made to improve the performance of contrastive learning by increasing batch size indirectly (He et al., 2020; Chen et al., 2020b). Therefore, to verify that the proposed mutual learning rather than the indirectly boosted batch sizes works, we double the batch size after removing mutual learning loss and refer it to *w/o* L_{mut} + *bsz*↑. The results show that despite the boosted batch size, it still performs worse than ML-LMCL, which demonstrate that the improvements come from the proposed mutual language rather than the boosted batch size.

3.5.2 Effectiveness of Distance Polarization Regularizer

To verify the effectiveness of distance polarization regularizer, we also remove distance polarization regularizer in pre-training and fine-tuning, which is named as *w/o* $\mathcal{L}_{reg}$ and *w/o* $\mathcal{L}^{p}_{reg}$ & $\mathcal{L}^{p}_{reg}$, respectively. When $\mathcal{L}_{reg}$ is removed, the accuracy drops by 0.24, 0.23 and 0.19 on SLURP, ATIS and TREC6, respectively. And when $\mathcal{L}^{p}_{reg}$ and $\mathcal{L}^{q}_{reg}$ are removed, the accuracy drops by 0.41, 0.29 and 0.22 on SLURP, ATIS and TREC6. The results demonstrate that distance polarization regularizer can alleviate the negative impact of conventional contrastive learning. Furthermore, the drop in accuracy is greater when fine-tuning than when pre-training. We believe that the reason is that supervised contrast learning in fine-tuning is easier to be affected by label noise than unsupervised contrast learning in pre-training. As a result, more semantically similar pairs are incorrectly pushed away in fine-tuning when the regularizer is removed.

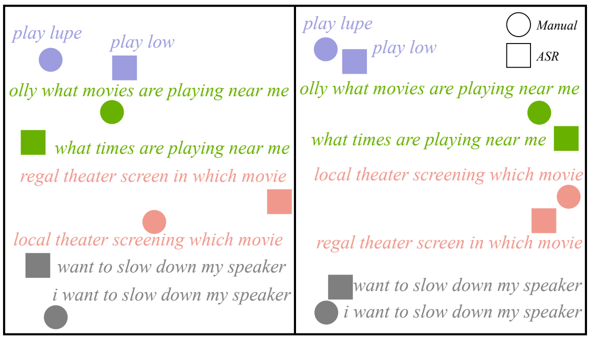
Chang and Chen (2022) also proposes a self-distilled soft contrastive learning loss to relieve the negative effect of noisy labels in supervised contrastive learning. However, we believe that the regularizer can also effectively reduce the impact of label noise. Therefore, our ML-LMCL does not include another module to tackle the problem of label noise. To verify this, we augument ML-LMCL with the self-distilled soft contrastive learning loss, which is termed as *w/* $\mathcal{L}_{soft}$. We can observe that not only $\mathcal{L}_{soft}$ does not bring any improvement, it even causes performance drops, which proves that the distance polarization regularizer can indeed reduce the impact of label noise.

3.5.3 Effectiveness of Cyclical Annealing Schedule

We also remove cyclical annealing schedule and relate it to *w/o cyc*. We observe that the accuracy drops by 0.18, 0.13 and 0.11 on SLURP, ATIS and TREC6, respectively, which demonstrates that the cyclical annealing schedule also plays an important role in enhancing the performance by mitigating the problem of KL vanishing.

3.6 Visualization

To better understand how mutual learning and large-margin contrastive learning affects and contributes to the final result, we show the visualization of an example on SLURP dataset in Figure 4. *“local theater screening which movie”* and *“olly what movies are playing near me”* are two manual transcripts with the same intent, and the representations of them and their associated ASR transcripts stay close to each other in ML-LMCL. However, in SpokenCSE, their representations keep a longer dis-



(a) SpokenCSE (b) ML-LMCL

Figure 4: Visualization of representations of manual transcripts and ASR transcripts. We visualize the representations by reducing the dimension with Principal Component Analysis (PCA) (Abdi and Williams, 2010). The circle and square in the same color means the corresponding manual and ASR transcriptions are associated.

tance, which further demonstrates that our method can align the ASR transcript and its associate manual transcript with high accuracy and better avoid semantically similar pairs being pushed away.

4 Related work

Error-robust Spoken Language Understanding SLU usually suffers from ASR error propagation and this paper focus on improving ASR robustness in SLU. Chang and Chen (2022) makes the first attempt to use contrastive learning to improve ASR robustness with only textual information. Following Chang and Chen (2022), this paper only focuses on intent detection in SLU. Intent detection is usually formulated as an utterance classification problem. As a large number of pre-trained models achieve surprising results across various tasks (Dong et al., 2022; Yang et al., 2023c; Zhu et al., 2023; Yang et al., 2023b), some BERT-based (Devlin et al., 2019) pre-trained work has been explored in SLU where the representation of the special token [CLS] is used for intent detection. In our work, we adopt RoBERTa and try to learn the invariant representations between clean manual transcripts and erroneous ASR transcripts.

Mutual Learning Our method is motivated by the recent success in mutual learning. Mutual learning is an effective method which trains two models of the same architecture simultaneously but with different initialization and encourages them to learn collaboratively from each other. Unlike knowledge distillation (Hinton et al., 2015), mutual learning doesn't need a powerful teacher network which is not always available. Mutual learning is first proposed to leverage information from multiple models and allow effective dual knowledge transfer in image processing tasks (Zhang et al., 2018; Zhao et al., 2021). Based on this, Wu et al. (2019b) utilizes mutual learning to capture complementary features in semi-supervised classification. Wu et al. (2019a) applies mutual learning between contour extraction and edge extraction for saliency detection. In NLP, Zhao et al. (2021) utilizes mutual learning for speech translation to transfer knowledge between a speech translation model and a machine translation model. In our work, we apply a mutual learning framework to transfer knowledge between the model trained on manual transcripts and the model trained on ASR transcripts.

Contrastive learning Contrastive learning aims at learning example representations by minimizing the distance between the positive pairs in the vector space and maximizing the distance between the negative pairs (Saunshi et al., 2019; Liang et al., 2022; Liu et al., 2022a), which is first proposed in the field of computer vision (Chopra et al., 2005; Schroff et al., 2015; Sohn, 2016; Chen et al., 2020a; Wang and Liu, 2021). In the NLP area, contrastive learning is applied to learn sentence embeddings (Giorgi et al., 2021; Yan et al., 2021), translation (Pan et al., 2021; Ye et al., 2022) and summarization (Wang et al., 2021; Cao and Wang, 2021). Recently, Chen et al. (2021) points that conventional contrastive learning algorithms are still not good enough since they fail to maintain a large margin in the distance space for reliable instance discrimination Inspired by this, we add a similar distance polarization regularizer as Chen et al. (2021) to address this issue. To the best of our knowledge, we are the first to introduce the idea of large-margin contrastive learning to the SLU task.

5 Conclusion

In this paper, we propose ML-LMCL, a novel framework for improving ASR robustness in SLU. We apply mutual learning and introduce the distance polarization regularizer. Moreover, cyclical annealing schedule is utilized to mitigate KL vanishing. Experiments and analysis on three benchmark datasets show that our model significantly outperforms previous models whether clean manual transcriptions is available in fine-tuning or not. Future work will focus on improving ASR robustness with only clean manual transcriptions.

Limitations

By applying mutual learning, introducing distance polarization regularizer and utilizing cyclical annealing schedule, ML-LMCL achieves significant improvement on three benchmark datasets. Nevertheless, we summarize two limitations for further discussion and investigation of other researchers:

(1) ML-LMCL still requires the ASR transcripts in fine-tuning to align with the target inference scenario. However, the ASR transcripts may not always be readily available due to the constraint of ASR systems and privacy concerns. In the future work, we will attempt to further improve ASR robustness without using any ASR transcripts.

(2) The training and inference runtime of ML-LMCL is larger than that of baselines. We attribute the extra cost to the fact that ML-LMCL has more parameters than baselines. In the future work, we plan to design a new paradigm with fewer parameters to reduce the requirement for GPU resources.

Acknowledgements

We thank all anonymous reviewers for their constructive comments. This paper was partially supported by Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001) and NSFC (No: 62176008).

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Ya-Hsin Chang and Yun-Nung Chen. 2022. Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding. In *Proc. Interspeech 2022*, pages 3458–3462.

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

Shuo Chen, Gang Niu, Chen Gong, Jun Li, Jian Yang, and Masashi Sugiyama. 2021. Large-margin contrastive learning with distance polarization regularizer. In *International Conference on Machine Learning*, pages 1673–1683. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023a. M3st: Mix at three levels for speech translation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2023b. Ssvmr: Saliency-based self-training for video-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Ganesh Ramakrishnan, and Preethi Jyothi. 2022. Error correction in asr using sequence-to-sequence models. *arXiv preprint arXiv:2202.01157*.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

E Haihong, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5467–5471.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Rian He, Shubin Cai, Zhong Ming, and Jialei Zhang. 2022. Weighted self distillation for chinese word segmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1757–1770.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. 2021. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522.

Chao-Wei Huang and Yun-Nung Chen. 2019. Adapting pretrained transformer to lattices for spoken language understanding. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 845–852. IEEE.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5577–5587.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3202–3213.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91.

Baohao Liao, Yingbo Gao, and Hermann Ney. 2020. Multi-agent mutual learning at sentence-level and token-level for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1715–1724.

Risheng Liu, Zhiying Jiang, Shuzhou Yang, and Xin Fan. 2022a. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31:4922–4936.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022b. Are we ready for a new paradigm shift? a survey on visual deep mlp. *Patterns*, 3(7):100520.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. Fastbert: a self-distilling bert with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6035–6044.

Yang Liu, Sheng Shen, and Mirella Lapata. 2021. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–703.

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 arXiv:1907.11692*.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE.

Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30.

Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 502–517.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Gwenaelle Cunha Sergio, Dennis Singh Moirangthem, and Minho Lee. 2020. Attentively embracing noise for robust latent representation in bert. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3479–3491.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript. *arXiv preprint arXiv:2102.00804*.

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

Chengyu Wang, Suyang Dai, Yipeng Wang, Fei Yang, Minghui Qiu, Kehan Chen, Wei Zhou, and Jun Huang. 2022. Arobert: An asr robust pre-trained language model for spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1207–1218.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Haoyu Wang, Shuyan Dong, Yue Liu, James Logan, Ashish Kumar Agrawal, and Yang Liu. 2020. Asr error correction with augmented transformer for entity retrieval. In *Interspeech*, pages 1550–1554.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. 2019a. A mutual learning method for salient object detection with intertwined multi-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8150–8159.

Si Wu, Jichang Li, Cheng Liu, Zhiwen Yu, and Hau-San Wong. 2019b. Mutual learning of complementary networks via residual correction for improving semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6500–6509.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023a. Multicapclip: Auto-encoding prompts for zero-shot multilingual visual captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Bang Yang, Fenglin Liu, Yuexian Zou, Xian Wu, Yaowei Wang, and David A Clifton. 2023b. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *arXiv preprint arXiv:2303.06458*.

Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. 2023c. Implicit neural representation for cooperative low-light image enhancement. *arXiv preprint arXiv:2303.11722*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5099–5113. Association for Computational Linguistics.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023b. Dub: Discrete unit back-translation for speech translation. *arXiv preprint arXiv:2305.11411*.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903.

Jiawei Zhao, Wei Luo, Boxing Chen, and Andrew Gilman. 2021. Mutual-learning improves end-to-end speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3989–3994.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141.

Wei Zhu, Xiaoling Wang, Yuan Ni, and Guotong Xie. 2021. Gaml-bert: Improving bert early exiting by gradient aligned mutual learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3033–3044.

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

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Limitation Section.

☒ A2. Did you discuss any potential risks of your work?
This paper does not involve any data collection and release thus there are no privacy issues. All the datasets used in this paper are publicly available and widely adopted by researchers to test the performance of SLU models.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Section Abstract and Section 1. Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In section 3. Experiments.

☑ B1. Did you cite the creators of artifacts you used?
In section 3. Experiments.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In section 3. Experiments.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In section 3. Experiments.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In section 3. Experiments.

C ☑ Did you run computational experiments?

In section 3. Experiments.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 3. Experiments.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In section 3. Experiments.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In section 3. Experiments.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In section 3. Experiments.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.