Zero-Shot End-to-End Spoken Language Understanding via Cross-Modal Selective Self-Training

Jianfeng He[1,2*], **Julian Salazar**[1], **Kaisheng Yao**[1], **Haoqi Li**[1], **Jinglun Cai**[1],
[1] AWS AI Labs, Seattle, Washington, USA
[2]Sanghani Center for Artificial Intelligence & Data Analytics, Virginia Tech, Falls Church, VA, USA
jianfenghe@vt.edu, {julsal, kaishey, haoqili, cjinglun}@amazon.com

Abstract

End-to-end (E2E) spoken language understanding (SLU) is constrained by the cost of collecting speech-semantics pairs, especially when label domains change. Hence, we explore *zero-shot* E2E SLU, which learns E2E SLU without speech-semantics pairs, instead using only speech-text and text-semantics pairs. Previous work achieved zero-shot by pseudolabeling all speech-text transcripts with a natural language understanding (NLU) model learned on text-semantics corpora. However, this method requires the domains of speech-text and text-semantics to match, which often mismatch due to separate collections. Furthermore, using the entire speech-text corpus from any domains leads to *imbalance* and *noise* issues. To address these, we propose *cross-modal selective self-training* (CMSST). CMSST tackles imbalance by clustering in a joint space of the three modalities (speech, text, and semantics) and handles label noise with a selection network. We also introduce two benchmarks for zero-shot E2E SLU, covering matched and found speech (mismatched) settings. Experiments show that CMSST improves performance in both two settings, with significantly reduced sample sizes and training time.

1 Introduction

End-to-end (E2E) spoken language understanding (SLU) models train on speech-semantics pairs, inferring semantics directly from acoustic features (Serdyuk et al., 2018) and leveraging non-lexical information like stress and intonation. In contrast, pipelined SLU models (Tur and De Mori, 2011) operate on speech-transcribed text, omitting the acoustic information. In all, E2E SLU has gained significant research attention. However, training E2E SLU models faces a significant challenge in collecting numerous speech-semantics pairs (Hsu et al., 2021). This challenge is two-fold: the scarcity of public speech-semantics pairs due to annotation costs and the need to relabel speeches when the labeling schema evolves, e.g., functionality expansion (Goyal et al., 2018). While speech-semantics pairs are scarce and expensive to annotate, there is a growing availability of speech-text pairs used in automatic speech recognition (ASR) and text-semantics pairs used in natural language understanding (NLU) (Galvez et al., 2021; FitzGerald et al., 2022). Thus, we define *zero-shot* E2E SLU, which learns an E2E SLU model by speech-text and text-semantics pairs *without ground-truth speech-semantics pairs* (hence zero-shot).

Only two works have explored zero-shot E2E SLU. Pasad et al. (2022) trained an NLU model by text-semantics pairs and used it to predict pseudolabels for the text of *all* speech-text pairs, similar to Figure 1(a). They then trained an E2E SLU model using the speech audio from the speech-text pairs, paired with the predicted pseudolabels. In another way, Mdhaffar et al. (2022) mapped the text of *all* text-semantics pairs to speech embeddings, creating "pseudospeech"-semantics pairs.

However, both works assume that text-semantics and speech-text pairs have matched domains. In practice, however, the speech-text and text-semantics pairs are often separately collected, so the domain of speech in speech-text pairs and text in text-semantics pairs may be mismatched. In such cases, directly using all speech-text and text-semantics pairs for zero-shot E2E SLU leads to two types of issues, which we classify as:

Noise. *Sample noise* comes from speech-text pairs whose transcripts (texts) are out-of-domain (OOD) for the NLU task. Passing all transcripts through NLU inference leads to inaccurate pseudolabels on the OOD data, impacting SLU learning. This exacerbates *label noise*, which refers to incorrect NLU model predictions that are then (wrongly) treated as pseudolabels; this issue is inherent to self-training

*Work done during an internship at AWS AI Labs.

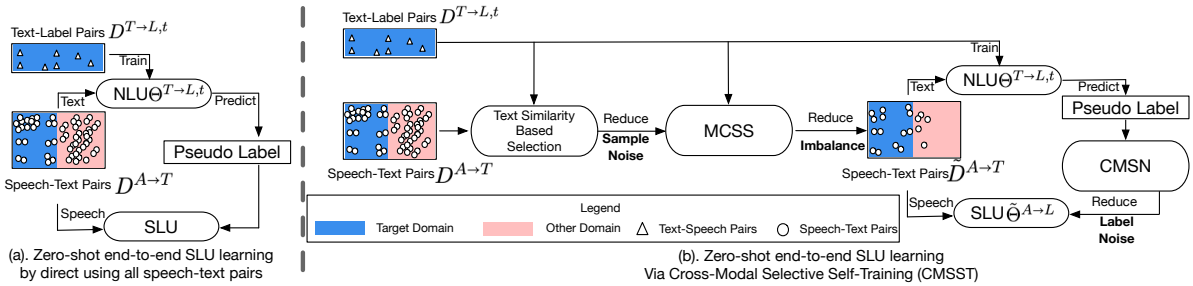

Figure 1: **(a)**. Diagram of using all speech-text pairs, detailed in Sec. 1. The legend in (b) is also applicable to (a). **(b)**. Diagram of the CMSST framework (described in Sec. 4). Speech and text pairs in $D^{A\to T}$ are selected by first using a text-similarity-based selection method and then a Multi-view Clustering-based Sample Selection (MCSS) algorithm. The SLU model $\tilde{\Theta}^{A\to L}$ is trained on the resulting speech-text pairs $\tilde{D}^{A\to T}$, with pseudolabels from an NLU model $\Theta^{T\to L,t}$. This NLU model is trained from target domain text-to-semantics pairs $D^{T\to L,t}$. To deal with label noise from the NLU model, CMSST uses a Cross-Modal SelectiveNet (CMSN) to train our SLU model $\tilde{\Theta}^{A\to L}$.

and also impacts performance (Du et al., 2020).

Imbalance. Since the text-semantics and speech-text pairs are separately collected, even after removing OOD speech-text pairs, the remaining text in speech-text pairs may be heavily imbalanced within the NLU domain, e.g., one semantics dominates all others. Besides, imbalanced speech, e.g., having only female voices, can bias E2E SLU learning. Though a model may succeed despite the imbalance, this can waste training resources that could have been used on representative speech-text pairs.

For these issues, Pasad et al. (2022) and Mdhaffar et al. (2022) mitigate sample noise and imbalance by selecting speech-text pairs that directly match the target text-semantics domain; however, in practice, it is hard to gain such well-matched and well-balanced speech-text corpus themselves. Furthermore, neither work is selective with pseudo-data, which in Pasad et al. (2022) led to degradation when more external speech-text was added, due to label noise. Instead, with *selection* as a unifying perspective, we make the following contributions:

(i). Zero-shot E2E SLU benchmarks for both matched and found speech. For the matched domain setting, we define **VoxPopuli2SLUE**, combining text-semantics pairs of SLUE's NER-annotated subset (Shon et al., 2022) of VoxPopuli (Wang et al., 2021) with speech-text pairs from VoxPopuli, similar to Pasad et al. (2022). More notably, for the found speech setting, we define **MiniPS2SLURP**, combining the home-assistant text-semantics pairs of SLURP (Bastianelli et al., 2020) with speech-text pairs from the general-domain People's Speech corpus (Galvez et al., 2021). The data split and code for our zero-shot E2E SLU benchmarks are publically released.[1]

(ii). Selection via cross-modal clustering and selective networks to tackle imbalance and noise in self-training. To tackle sample noise, we first exclude OOD speech-text pairs using text similarity. Then, for the imbalance, we propose *multi-view clustering-based sample selection (MCSS)* to resample speech-text pairs to give proportionate diversity over three views (speech, text and latent semantics). For label noise, we propose a *cross-modal SelectiveNet (CMSN)*, which selectively and learnably trusts pseudolabels based on the ease of learning common representations between the NLU and SLU encoders. All together, we refer to our proposed framework as **cross-modal selective self-training (CMSST)**, summarized in Figure 1(b).

(iii). Comprehensive experiments on zero-shot E2E SLU. We compare the baselines with our CMSST on the new benchmarks. CMSST achieves better results with significantly less data. Ablations show that clustering and selective learning both contribute; Entity F1 improves 1.2 points on VoxPopuli2SLUE with MCSS and 1.5 points on MiniPS2SLURP with CMSN.

2 Related Work

Speech to semantics. Although not fully zero-shot, works in semi-supervised E2E SLU have also considered the mismatch problem. Rao et al. (2020) train NLU and ASR systems independently, saving their task-specific SLU data for a final joint training stage. Others tackle the data sparsity or mismatch issues using text-to-speech (TTS) to synthesize spo-

[1] https://github.com/amazon-science/zero-shot-E2E-slu

Data	Annotation	MiniPS2 SLURP	VoxPopuli2 SLUE
$D^{A\to L,t}$	Speech-to-semantics pairs in target domain t	22,782	2,250
$D^{T\to L,t}$	Text-to-semantics pairs in target domain t	22,783	2,250
$D^{A\to T,t}$	Speech-to-text pairs in target domain t	22,782	2,250
$D^{A\to T,o}$	Speech-to-text pairs in other domains o	32,255	182,466
$D^{A\to T}$	Union of $D^{A\to T,t}$ and $D^{A\to T,o}$	55,037	184,716
Test	Test speech-to-semantics pairs in target domain t	13,078	877

Table 1: Data annotations and sample sizes in our datasets. $D^{A\to L,t}$ is used for training a target SLU model $\Theta^{A\to L,t}$. $D^{T\to L,t}$ and $D^{A\to T}$ are used for training our E2E SLU model $\tilde{\Theta}^{A\to L}$.

ken counterparts to NLU examples (Lugosch et al., 2020). Pretraining on off-the-shelf (found) speech-only data (Lugosch et al., 2019), text-only data (Huang et al., 2020), or both (Chung et al., 2020; Thomas et al., 2022) have improved SLU systems beyond their core speech-semantics training data, usually via an alignment objective or joint network. Finally, Rongali et al. (2021) considered a different notion of "zero-shot" E2E SLU, which we view more aptly as text-only SLU adaptation; their setting involves an initial E2E SLU model, trained on speech-semantics pairs, having its label set expanded with text-only data.

Self-training. This method (Scudder, 1965; Yarowsky, 1995) further trains a model on unlabeled inputs that are labeled by the same model, as a form of semi-supervised learning. It has experienced a recent revival in both ASR (Kahn et al., 2020) and NLU (Du et al., 2020), giving improvements atop strong supervised and self-supervised models, for which effective sample filters and label confidence models were key. Recently, Pasad et al. (2022) performed self-training in the zero-shot E2E NER case; however, since they work in the matched case they do not address these issues of imbalance and noise.

Multi-view clustering. Multiple views of the data can improve clustering by integrating extensive information (Kumar and Daumé, 2011; Wang et al., 2022; Qin et al., 2021; Xu et al., 2022). We propose using the modalities in speech-text pairs (speech, text, and latent semantics) as bases to build a joint space, where we apply clusters to enable balanced selection. We apply simple heuristics atop the clusters, and leave stronger algorithms, e.g., Trosten et al. (2021) to future work.

Selective learning. Selective learning aims at designing models that are robust in the presence of mislabeled datasets (Ziyin et al., 2020). It is often achieved by a selective function (Geifman and El-Yaniv, 2019). Selective learning has been recently applied in a variety of applications (Gangrade et al., 2021; Kühne and Gühmann, 2022; Varshney et al., 2022), but less so in NLP applications (Xin et al., 2021). To the best of our knowledge, we are the first to propose a selective learning method (Sec. 4.4) in the cross-modal setting.

3 Benchmarks for Zero-Shot E2E SLU

We define a target E2E SLU model as $\Theta^{A\to L,t}$, that is trained on data $D^{A\to L,t}$ with pairs of speech **audio** A and semantic **labels** L. These samples are in a target domain t. We also write superscript $T \to L$ to denote **text** T to semantic labels, and $A \to T$ to denote speech audio to text.

Hence, in zero-shot, instead of having a speech-to-semantics dataset $D^{A\to L,t}$, we have a text-to-semantics pair set $D^{T\to L,t}$ in the target domain, and an external speech-to-text pair set $D^{A\to T}$. Unlike Pasad et al. (2022) or Mdhaffar et al. (2022), the external speech-to-text data $D^{A\to T}$ may be independently collected and have sample pairs from other domains. We divide $D^{A\to T}$ into two disjoint subsets, with samples either in the **target domain** t or being **other domains** o:

$$D^{A\to T} = D^{A\to T,t} \cup D^{A\to T,o}. \quad (1)$$

Given $D^{T\to L,t}$ and $D^{A\to T}$, we aim to learn an E2E SLU model $\tilde{\Theta}^{A\to L}$ that performs close to $\Theta^{A\to L,t}$. This is zero-shot, as training our $\tilde{\Theta}^{A\to L}$ uses no speech-semantics pairs $D^{A\to L,t}$. We created the below two datasets to study this problem:

Matched Speech: VoxPopuli2SLUE. We use *SLUE-VoxPopuli* (Shon et al., 2022) as the target domain text-to-semantics data $D^{T\to L,t}$. The external speech-to-text data $D^{A\to T}$ is from *VoxPopuli* (Wang et al., 2021). Because the two are matched (SLUE-VoxPopuli and VoxPopuli are both from European Parliamentary proceedings). We denote this dataset as VoxPopuli2SLUE. In our notation, **matched** means

$$D^{A\to T} = D^{A\to T,t}, \quad D^{A\to T,o} = \emptyset. \quad (2)$$

Found Speech: MiniPS2SLURP. We use *SLURP* (Bastianelli et al., 2020) as the target domain text-to-semantics data $D^{T\to L,t}$. *MiniPS* (Galvez et al., 2021) provides the other-domain

speech-to-text pairs $D^{A\to T,o}$. SLURP is in the voice commands domain for controlling family robots. It consists of 18 scenarios, with semantics of "scenario", "action", "intent" and "entity". But Mini-PS is a subset of People's Speech corpus, with 32,255 speech-to-text pairs in diverse domains. We then mix $D^{A\to T,o}$ from Mini-PS and $D^{A\to T,t}$ from SLURP. The resulting dataset, MiniPS2SLURP, gives the **found** (mismatched) setting.

For fair comparison, in the above two datasets, we provide $D^{A\to L,t}$ that has the same size and speech as $D^{A\to T,t}$. The $D^{A\to L,t}$ is only used to learn $\Theta^{A\to L,t}$ and not applied to learn our $\tilde{\Theta}^{A\to L}$.

We use the full SLURP test set as the test set in MiniPS2SLURP, and half of the dev set in SLUE-VoxPopuli as the test set in VoxPopuli2SLUE. The dataset statistics, data annotations, and data usages are in Table 1 with sample data in Table 5.

4 Cross-Modal Selective Self-Training

4.1 Introduction of A Basic SLU Model

Given a sequence of acoustic features $\mathbf{A}$, the SLU models $\Theta^{A\to L,t}$ and $\tilde{\Theta}^{A\to L}$ extract sentence-level semantics (i.e., intents) and token-level semantics (i.e., entity tags). To support these multiple types of semantic tags, we use a sequence-to-sequence architecture (Bastianelli et al., 2020; Ravanelli et al., 2021), in which the output is a sequence $\mathbf{Y}$ that consists of semantic types with their tags. The SLU model uses a speech encoder to encode $\mathbf{A}$ into a sequence of speech representations, and uses an attentional sequence decoder to generate the output sequence $\mathbf{Y}$. The $\Theta^{A\to L,t}$ is trained by loss $\mathcal{L}^{A\to L}$ that maximizes the likelihood of generating the correct semantic sequence given the observation.

4.2 Overview of Our Model: CMSST

The speech-to-text data $D^{A\to T}$ could provide more external resource for SLU training. However, the domain mismatch across $D^{T\to L,t}$ and $D^{A\to T,o}$ is often inevitable and can lead to SLU performance loss. In addition, the large size of $D^{A\to T}$ may lead to inefficient model training. Thus, we propose a Cross-Modal Selective Self-Training (CMSST) framework to alleviate this domain mismatch impact in using $D^{A\to T}$ to learn the SLU model $\tilde{\Theta}^{A\to L}$. We later show in Table 2 that CMSST achieves higher performance and efficiency with fewer training samples in comparison to baselines.

Figure 1(b) illustrates CMSST. First, it computes text similarity to exclude samples in $D^{A\to T}$ with large divergence to $D^{T\to L,t}$. Second, it takes the distribution of the dataset into consideration, and further filters $D^{A\to T}$ using clustering methods to reduce the imbalance within $D^{A\to T}$ itself. These two steps are described in Sec. 4.3. Lastly, it uses a novel cross-modal selective training method, described in Sec. 4.4, to reduce the impact of noisy labels predicted by an NLU model $\Theta^{T\to L,t}$. The NLU model $\Theta^{T\to L,t}$ is pretrained on $D^{T\to L,t}$.

4.3 Reducing Sample Noise and Imbalance

Text similarity based selection. The sample selection is firstly performed in a text embedding space. K-means (Xu and Wunsch, 2005) is further employed to cluster in the text embedding space for texts from $D^{T\to L,t}$. For each text in $D^{A\to T}$, a text similarity score is defined as the distance to the closest clustering centroid of $D^{T\to L,t}$. Then a threshold based on the text similarity scores is set to exclude $D^{A\to T}$ pairs with text disparity.

Multi-view Clustering-based Sample Selection (MCSS). Though the above selection process removes speech-text pairs in the other domains, the remaining pairs can still be imbalanced. The imbalanced data distribution introduces bias into the training and decreases training efficiency. Therefore, it is important to balance the remaining speech-text pairs. Since each speech-text pair contains audio, text, and latent semantic information, we propose MCSS to balance these three components. Figure 2 illustrates MCSS's workflow. We use superscripts T, A, and L to each denote the text, speech, and semantic modalities, respectively.

First, for the text and speech modalities, we use K-Means to cluster samples in $D^{A\to T}$. The text embedding is SentenceBERT (Reimers and Gurevych, 2019) or the average of GloVe word2vec (Pennington et al., 2014). The speech embedding is the average of a low-layer feature map in HuBERT (Hsu et al., 2021). This step respectively outputs K^T and K^A numbers of clustering centroids of text modality and speech modality in $D^{A\to T}$.

To represent the semantic space, each entity type in $D^{T\to L,t}$ is an averaged text embedding on all text spans inside that entity type. Therefore, the number of entity centroids K^L is the number of entity types. We denote these centroids as $\{\mu_k^v\}$ for $k \in K^v$ and $v \in \{T, A, L\}$ across three modalities.

Given a sample $\mathbf{X}_i$ in $D^{A\to T}$, its distance to k-th clustering centroids μ_k^v in modality v is denoted as $d^v(\mathbf{X}_i, \mu_k^v)$. Then, we compute the sam-

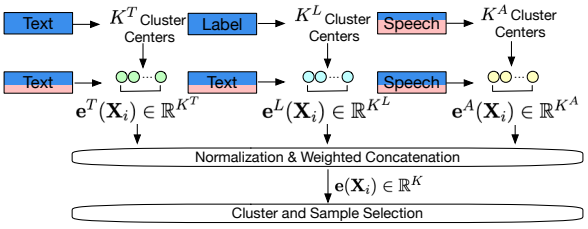

Figure 2: Diagram of MCSS (described in Sec. 4.3). We use superscripts T, A, and L to each denote text, speech, and semantic modality. A blue box represents the data from the $D^{T\to L,t}$, and a blue and pink box represents data from $D^{A\to T}$.

ple modality-specific view $e^v(\mathbf{X}_i) \in \mathbb{R}^{K^v}$ as the sample distances to all centroids in modality v,

$$e^v(\mathbf{X}_i) = [\cdots, d^v(\mathbf{X}_i, \mu_k^v), \cdots] \tag{3}$$

and $k \in \{1, 2, ..., K^v\}$.

Among three views, $\mathbf{e}^T(\mathbf{X}_i)$ and $\mathbf{e}^L(\mathbf{X}_i)$ contain information related to $T \to L$ domain, while $\mathbf{e}^A(\mathbf{X}_i)$ is generated from speech representation that highly correlates acoustic features in $D^{A\to T}$.

We use Cosine distance for the speech and text views and Mahalanobis distance for the semantic view. As they are in different scales, we apply zero-score normalization in each view. In addition, to address the different importance across different views, we use adjustable scalar weight for each view. The multi-view representation is then created by weighted concatenations as $\mathbf{e}(\mathbf{X}_i) = [w^T\mathbf{e}^T(\mathbf{X}_i), w^A\mathbf{e}^A(\mathbf{X}_i), w^L\mathbf{e}^L(\mathbf{X}_i)]$ and $\mathbf{e}(\mathbf{X}_i) \in \mathbb{R}^K$ with $K = K^T + K^A + K^L$.

We again apply the K-Means algorithm on these multi-view representations $\{\mathbf{e}(\mathbf{X}_i)\}$ by setting R clusters. The corresponding clusters represent "supports" of a joint text, speech, and semantics space.

To obtain samples that are balanced in this joint space, we select the equal number of samples for each cluster, and these samples are nearest to the cluster centroid they belong to. Suppose we target for N samples out of the algorithm, then each cluster selects ($\lfloor \frac{N}{R} \rfloor$) of the nearest samples. More details are in Sec. A.1.

4.4 Reducing Label Noise

Given the selected speech-to-text pair set $\tilde{D}^{A\to T}$ from MCSS, the pretrained NLU model $\Theta^{T\to L,t}$ predicts pseudolabels. An SLU model is then trained on the speech and its pseudolabels. However, these pseudolabels are noisy, because of prediction errors in the imperfect NLU model $\Theta^{T\to L,t}$. Hence, we propose **Cross-Modal SelectiveNet (CMSN)** to support selective learning and reduce the label noise.

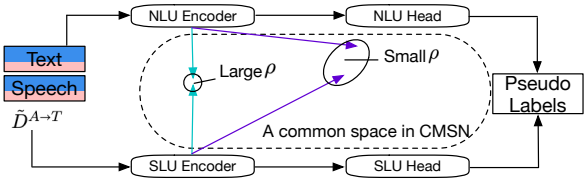

Figure 3: Diagram of workflow for CMSN (described in Sec. 4.4), where green or purple arrows are a pair of text and speech. ρ is a selective score described in Eq. (6).

Figure 3 illustrates our CMSN. For a speech-to-text pair $\mathbf{X}_i$ from $\tilde{D}^{A\to T}$, a text encoder in $\Theta^{T\to L,t}$ and a speech encoder in $\tilde{\Theta}^{A\to L}$ extract their modality-specific embedding vector $\mathbf{f}_i^T$ and $\mathbf{f}_i^A$. Because these embeddings are from the same speech-to-text pair in $\tilde{D}^{A\to T}$, they share a common semantic space. Therefore, we learn modality-specific projections to map the i-th sample embeddings to vectors with the same shapes as below,

$$\mathbf{p}_i^v = \mathbf{P}^v\mathbf{f}_i^v,\ \mathbf{q}_i^v = \mathbf{Q}^v\mathbf{f}_i^v \tag{4}$$

where $v \in \{T, A\}$ and $\mathbf{q}$ is from the second common space introduced later. We can measure cross-modal loss $\mathcal{L}_{cm1_i}$ by the divergence between their common semantic space representations,

$$\mathcal{L}_{cm1_i} = ||\mathbf{p}_i^T - \mathbf{p}_i^A|| \tag{5}$$

To facilitate selective learning, we compute a scalar selective score $\rho \in (0, 1)$ through a selection function $g(\cdot)$ as below,

$$\rho_i = g(\mathbf{p}_i^T, \mathbf{p}_i^A) \tag{6}$$

g is a multilayer perceptron with a sigmoid function on top of the last layer. With the selective score, we define the following selective learning loss $\mathcal{L}_{sel}$ to abstain samples with low selection scores,

$$\begin{aligned}\mathcal{L}_{sel} &= \alpha \cdot [max(\tau - E[\rho_i], 0)]^2 \\ &+ \beta \cdot \frac{E[\rho_i\mathcal{L}_{cm1_i} + \rho_i\mathcal{L}^{A\to L}]}{E[\rho_i]}\end{aligned} \tag{7}$$

where α and β are scalar weights. The first term in Eq. (7) has a hyper-parameter $\tau \in [0, 1]$, which is defined as the target coverage in Geifman and El-Yaniv (2019). The first term encourages the selective network to output selective scores that are approaching τ, especially if the selective scores are small at the beginning of model training.

For the second term in Eq. (7), when this loss $\mathcal{L}_{cm1_i}$ is large because of a biased text embedding, the second term encourages a small ρ_i for Eq. (6) to learn. Due to the biased text embedding, its respective pseudolabel is also biased. Thus, even if a biased pseudolabel leads to large $\mathcal{L}^{A\to L}$, its impact is scaled down by ρ_i. The final loss is below,

$$\mathcal{L} = \mathcal{L}^{A\to L} + \mathcal{L}_{sel} + \gamma\mathcal{L}_{cm_2} \tag{8}$$

where γ is the weight of auxiliary cross-modal loss $\mathcal{L}_{cm_2}$. The $\mathcal{L}_{cm_2}$ encourages the common space learning by the expectation (mean) of all sample cross-modal differences weighted by respective ρ,

$$\mathcal{L}_{cm2} = E[\rho_i||\mathbf{q}_i^T - \mathbf{q}_i^A||] \tag{9}$$

The use of the $\mathcal{L}_{cm2}$ via another projection $\mathbf{Q}^v$ is essential to optimize selective network (Geifman and El-Yaniv, 2019). With $\mathcal{L}_{cm2}$, the selective network can additionally learn the alignment of cross-modal features. Therefore, $\mathcal{L}_{cm2}$ avoids overfitting the selective network to the wrong subset, before accurate low-level speech features are learned.

5 Experiments

We now compare models trained with the proposed framework with alternative models on the two zero-shot SLU datasets introduced in Sec. 3.

5.1 Performance Metrics

Following (Bastianelli et al., 2020), we report 1) sentence-level classification performance using average accuracy (**Avg. Acc.**) on classifying Scenario (Scenario Acc.), action (Action Acc.) and intent (Intent Acc.), and 2) NER performance from the list of entity type-value pairs. The **Entity-F1** is a sentence-level NER metric, in which the correctness of entity type-value pairs and their appearance orders are measured. **Word-F1** drops the penalty on their appearance orders. **Char-F1** further relaxes exact match at word level and allows character-level match of entity values. To measure the training efficiency, we report numbers of used speech-text pairs (sum of $\|D^{A\to T,t}\|$ and $\|D^{A\to T,o}\|$) and training time. Experiments were run on a single GPU 3090 with 24G memory.

5.2 Baselines & Experiment Setups

We compare our method with two types of methods: 1) a strong baseline that uses all of the ASR data (Pasad et al., 2022), denoted as $\tilde{\Theta}^{A\to L}_{Full}$ and 2) a model that random samples training data to have data size comparable to our method, denoted as $\tilde{\Theta}^{A\to L}_{RSamp}$. We also report the performance of $\Theta^{A\to L,t}$ that is trained with target domain speech-to-semantics data $D^{A\to L,t}$. We compare text-similarity selection by GloVe and Sentence-BERT(Abbr: SentBERT) . The ablation studies are GloVe-based.

5.3 Main Results

The main results of the proposed model on the two datasets are illustrated in Table 2. Firstly, our proposed method using SentBERT embedding can surpass the strong baseline $\tilde{\Theta}^{A\to L}_{Full}$ that uses all training samples in both GloVe-based and SentBERT-based text-similarity. For example, on the NER task, our SentBERT-based model has entity-F1 on the matched speech VoxPopuli2SLUE is 38.0%, surpassing the full system that is 37.0%. Besides, our method shows a significant reduction of training time from 225 hours to 6 hours and number of speech-text pairs from 18k to 5k, as our method uses 3% of the full dataset size. On the found speech MiniPS2SLURP, our SentBERT-based model achieves higher performance in both accuracy and F1 scores and higher training efficiency. For example, it improves 1.2 points in Entity F1 than $\tilde{\Theta}^{A\to L}_{Full}$ that uses 1.5 times of training time and data size of ours.

Our performance gain is apparent when compared to$\tilde{\Theta}^{A\to L}_{RSamp}$, using a similar size of randomly sampled training data. In such a case, entity F1 scores on two datasets drop by around 1 and 2 percents compared to our GloVe-based and SentBERT-based methods, respectively.

The proposed method surpasses the performance of the target model $\Theta^{A\to L,t}$ in the matched speech VoxPopuli2SLUE set. For instance, our SentBERT-based model has word-level entity F1 improved to 49.3% from 45.2% of the target model. On the found speech MiniPS2SLURP, the difference to the target model is reduced to 0.6% by our method, compared to 1.1% by $\tilde{\Theta}^{A\to L}_{Full}$ and 2.5% by $\tilde{\Theta}^{A\to L}_{RSamp}$ in terms of Avg. Acc.

The results on SentBERT-based text-similarity marginally perform better than the GloVe-based. Except the 1.2 percents difference on NER F1 on VoxPopuli2SLUE, all the other metrics on both two datasets show less than 1 percent difference. The marginal difference between two methods is similar to other self-training work (Du et al., 2020).

Models	$\|D^{A\to L,t}\|$	$\|D^{A\to T,t}\|$	$\|D^{A\to T,o}\|$	Avg. Acc. (in %)	NER F1 (in %) Entity	Word	Char	Time (in hrs)
MiniPS2SLURP								
Target model $\Theta^{A\to L,t}$	22.8k	0	0	76.0	40.9	51.7	55.8	16
$\tilde{\Theta}^{A\to L}_{Full}$ (Pasad et al., 2022)	0	22.8k	32.3k	74.9	34.9	48.8	52.0	43
$\tilde{\Theta}^{A\to L}_{RSamp}$	0	14.3k	20.6k	73.5	33.9	47.5	50.9	27
Our model $\tilde{\Theta}^{A\to L}$ (GloVe)	0	21.6k	13.4k	75.2	34.9	48.8	52.2	28
Our model $\tilde{\Theta}^{A\to L}$ (SentBERT)	0	22.1k	12.9k	**75.4**	**35.7**	**49.3**	**52.9**	27
VoxPopuli2SLUE								
Target model $\Theta^{A\to L,t}$	2,250	0	0	N/A	36.0	45.2	47.7	2
$\tilde{\Theta}^{A\to L}_{Full}$ (Pasad et al., 2022)	0	2,250	182.5k	N/A	37.0	**50.3**	**53.9**	225
$\tilde{\Theta}^{A\to L}_{RSamp}$	0	68	5.6k	N/A	35.7	47.8	50.5	6
Our model $\tilde{\Theta}^{A\to L}$ (GloVe)	0	59	5.5k	N/A	36.8	49.0	52.3	6
Our model $\tilde{\Theta}^{A\to L}$ (SentBERT)	0	61	5.5k	N/A	**38.0**	49.3	52.4	6

Table 2: Results of the proposed CMSST and baselines on the datasets. Our model uses much less number of speech-text pairs (the sum of $\|D^{A\to T,t}\|$ and $\|D^{A\to T,o}\|$) and training time compared with using all speech-text pairs (Pasad et al., 2022), but achieves better or similar accuracy and F1 scores.

Due to the slight difference, our ablation studies use GloVe-based text similarity selection for faster speed.

6 Analysis

6.1 Ablation Studies

Multi-view Clustering-based Sample Selection(MCSS). We use different thresholds on the text similarity scores and control the final selected ASR data size with and without MCSS to be approximately the same for a fair comparison. Results are shown in Figure 4. On the found speech MiniPS2SLURP, we observe that removing MCSS (**w/o MCSS**) hurts performance. For example, using MCSS, entity F1 score is improved from 18.8% to 28.0%, a 49% relative improvement. Another observation is that MCSS apparently has more in-domain samples than without using the MCSS algorithm. For instance, the number of out-of-domain samples is 10350 and is almost twice the samples selected via MCSS in $\tilde{\Theta}^{A\to L}$.

Cross Modal SelectiveNet (CMSN). Results in Figure 4 show that further removing selective training (w/o MCSS, w/o CMSN) results in performance loss. On the MiniPS2SLURP, the entity F1 score is improved from 17.3% to 18.8% if using CMSN, a relative 8.7% improvement.

Performance improvements are also observed for the matched speech VoxPopuli2SLUE dataset in Figure 4. These results show that both reducing imbalance by sample selection (MCSS) and reducing label noise by selective learning (CMSN) contribute to the improved performance of the proposed framework.

6.2 Impacts from NLU Backbone

Backbone	MCSS+CMSN	NER F1 (in %) Entity	Word	Char
LSTM		35.1	45.5	48.6
	✓	**36.6**	**46.4**	**49.1**
BERT		35.0	47.3	50.4
	✓	**36.8**	**49.0**	**52.3**

Table 3: Impact comparison of using LSTM and BERT NLU backbones, on VoxPopuli2SLUE. Both backbones have $\|D^{A\to T,t}\| = 68$ and $\|D^{A\to T,o}\| = 5489$ after text similarity based selection and MCSS.

In this section, we conduct experiments on VoxPopuli2SLUE to study the impact of different NLU backbones in $\Theta^{T\to L,t}$. The comparison reveals the effectiveness of the proposed framework in dealing with different qualities of pseudolabels. We select LSTM and BERT due to their wide applications. The BERT-based backbone was fine-tuned from pretrained "bert-base-uncased". We fix its parameters but train prediction heads. The LSTM backbone was trained from scratch. Both backbones are trained from 2250 samples in $D^{T\to L,t}$. We measure their performance on the test set using ground truths from their text inputs. The BERT-based NLU backbone has higher NER performance than the LSTM-based NLU backbone, with 39.3% vs. 36.7% entity F1 Score.

From Table 3, we observe that (1) labels from BERT-based backbone result in comparable or higher performance, (2) using the framework (w/ MCSS+CMSN checked) consistently improves performances of the learned SLU models.

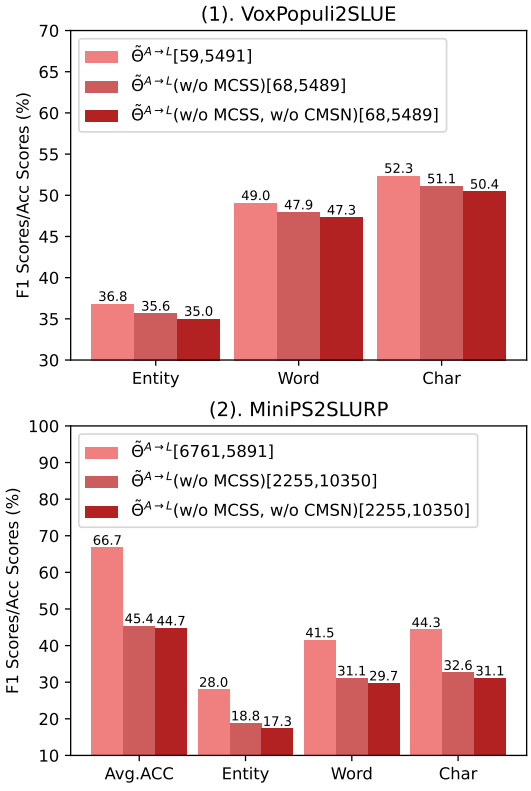

Figure 4: Ablation study on the effectiveness of multi-view sample selection and selective training on $\tilde{\Theta}^{A\to L}$. The pseudolabels are from BERT-based $\Theta^{T\to L,t}$. Their $\|D^{A\to T,t}\|$ and $\|D^{A\to T,o}\|$ size are each listed in square brackets for each configuration.

6.3 Sample Diversity

This section provides further analysis of MCSS. The observation in Figure 4 shows improved performance and increased proportions of in-domain data. Our hypothesis is that samples are more diverse due to the sample selection method described in Sec. 4.3. To quantify this, we measure the entropy of the selected samples, specifically for each view $v \in \{T, L, A\}$. Entropy in each view v is computed as $-\sum_{k=1}^{K^v} \frac{n_k^v}{N} \log \frac{n_k^v}{N}$, where K^v is the number of clusters for view v, n_k^v is the number of samples in cluster k for view v, and N is the total

Sampling Method	$\|D^{A\to T,t}\|$	$\|D^{A\to T,o}\|$	Diversity (Entropy) T	L	A
Equal	59	5,491	**3.94**	**1.34**	**4.36**
Random	61	5,495	3.84	1.24	4.34
Extreme	47	5,509	3.78	1.20	2.55
w/o MCSS	68	5,489	2.75	1.03	4.27

Table 4: Sample diversity from views of the three modalities (text (T), semantic labels (L), and audio (A)). They are computed as entropy on samples from different selection methods. Results are on VoxPopuli2SLUE.

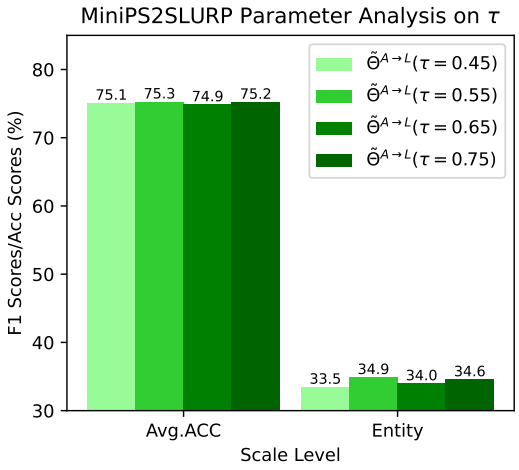

Figure 5: Entity F1 Scores and Avg. Acc. on the found speech MiniPS2SLURP dataset, where all groups have the same $\|D^{A\to T,t}\| = 21597$ and $\|D^{A\to T,o}\| = 13400$.

number of samples. Their results are in Table 4. For comparison, we also measure the entropy from random sampling (Random) and entropy from selecting samples with as few clusters as possible (Extreme). We observe that the entropy from the equal sampling method is larger than random sampling in all three views. The extreme sampling method has the lowest entropy, compared to the other two sampling methods. As a larger entropy indicates more diversity, we conclude that our equal sampling results in the largest diversity among these methods. We also list the entropy on a similar size of filtered samples without MCSS; their entropies in three views are much lower compared to our equal sampling method.

6.4 Parameter Analysis

Figure 5 shows Entity F1 scores and average accuracy on the found speech MiniPS2SLURP dataset. The pseudolabels are from the BERT-based $\Theta^{T\to L,t}$. We observe a performance dependency on the coverage rate τ with an optimal value of $\tau = 0.55$. Other parameter analysis reusults in both MCSS and CMSN are in Sec. A.6.

Case study of our model is in Table 6.

7 Conclusion

We have presented a method for zero-shot E2E spoken language understanding. We designed the method with the assumption that 1) speech-to-text and text-to-semantics data are collected separately and 2) speech-to-semantics data for an SLU model is not available. This is a challenging situation that

often happens in developing E2E SLU models for new applications and in new domains. To support this study, we have created two datasets: one for matched speech to the target domain and the other for found speech in diverse source domains. We have proposed methods to address two particular issues: 1) noise, which includes input noise from out-of-domain ASR data whose text transcripts are outliers in the NLU data domain and labeling noise from imperfect NLU models, and 2) imbalance, which often occurs in the multiple modalities of speech, text, and semantics for SLU. We have proposed a multi-view clustering-based sample selection method to select speech-text pairs that are representative of acoustic variability, text variability, and semantic coverage, aiming at reducing the imbalance. We further proposed a selective training model, Cross-Modal SelectiveNet, that attenuates the impact of low-confidence pseudolabels, aiming to reduce impacts from label noise. Extensive experiments on both datasets show that our methods achieve consistent improvement, approaching targeted direct E2E SLU models at a much lower computational cost than alternatives.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Yu-An Chung, Chenguang Zhu, and Michael Zeng. 2020. Splat: Speech-language joint pre-training for spoken language understanding. *arXiv preprint arXiv:2010.02295*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint arXiv:2010.02194*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gökhan Tür, and Prem Natarajan. 2022. MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *CoRR*, abs/2204.08582.

Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Felipe Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *NeurIPS Datasets and Benchmarks*.

Aditya Gangrade, Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. 2021. Online selective classification with limited feedback. *Advances in Neural Information Processing Systems*, 34:14529–14541.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR.

Anuj Goyal, Angeliki Metallinou, and Spyros Matsoukas. 2018. Fast and scalable expansion of natural language understanding functionality for intelligent agents. *arXiv preprint arXiv:1805.01542*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yinghui Huang, Hong-Kwang Kuo, Samuel Thomas, Zvi Kons, Kartik Audhkhasi, Brian Kingsbury, Ron Hoory, and Michael Picheny. 2020. Leveraging unpaired text data for training end-to-end speech-to-intent systems. In *ICASSP*, pages 7984–7988. IEEE.

Jacob Kahn, Ann Lee, and Awni Y. Hannun. 2020. Self-training for end-to-end speech recognition. In *ICASSP*, pages 7084–7088. IEEE.

Joana Kühne and Clemens Gühmann. 2022. Defending against adversarial attacks on time-series with selective classification. In *2022 Prognostics and Health Management Conference (PHM-2022 London)*, pages 169–175. IEEE.

Abhishek Kumar and Hal Daumé. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 393–400. Citeseer.

Loren Lugosch, Brett Meyer, Derek Nowrouzezaharai, and Mirco Ravanelli. 2020. Using speech synthesis to train end-to-end spoken language understanding models. In *ICASSP*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. In *INTERSPEECH*, pages 814–818. ISCA.

Salima Mdhaffar, Jarod Duret, Titouan Parcollet, and Yannick Estève. 2022. End-to-end model for named entity recognition from speech without paired training data. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 4068–4072. ISCA.

Ankita Pasad, Felix Wu, Suwon Shon, Karen Livescu, and Kyu Han. 2022. On the use of external data for spoken named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 724–737, Seattle, United States. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Yalan Qin, Hanzhou Wu, Xinpeng Zhang, and Guorui Feng. 2021. Semi-supervised structured subspace learning for multi-view clustering. *IEEE Transactions on Image Processing*, 31:1–14.

Milind Rao, Anirudh Raju, Pranav Dheram, Bach Bui, and Ariya Rastrow. 2020. Speech to semantics: Improve asr and nlu jointly via all-neural interfaces. In *INTERSPPECH*, pages 876–880.

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 arXiv:2106.04624*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Subendhu Rongali, Beiye Liu, Liwei Cai, Konstantine Arkoudas, Chengwei Su, and Wael Hamza. 2021. Exploring transfer learning for end-to-end spoken language understanding. In *AAAI*, pages 13754–13761. AAAI Press.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

Samuel Thomas, Hong-Kwang Jeff Kuo, Brian Kingsbury, and George Saon. 2022. Towards reducing the need for speech training data to build spoken language understanding systems. In *ICASSP*, pages 7932–7936. IEEE.

Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1255–1265.

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

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. *arXiv preprint arXiv:2203.00211*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. 2022. Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9776–9785.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. 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 1040–1051.

Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, S Yu Philip, and Lifang He. 2022. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Rui Xu and Donald Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196. Morgan Kaufmann Publishers / ACL.

Liu Ziyin, Blair Chen, Ru Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2020. Learning not to learn in the presence of noisy labels. *arXiv preprint arXiv:2002.06541*.

A Appendix

A.1 Model

Semantic representations. Specifically, the semantics in $D^{T\to L,t}$ has K^L types (i.e. "LOC", "DATE"). We use the average GloVe word2vec features of all slot texts from a semantic type to build their type centroids. As a result, we have K^L clustering centroids for semantics.

Normalization methods. For the normalization, we use the z-score normalization for $e^v(\mathbf{X}_i)$, where $v \in \{T, A, L\}$. After the normalization, each single-view representation $e^v(\mathbf{X}_i)$ obeys a standard Gaussian distribution and becomes comparable due to the same scale.

Special cases in selecting $\lfloor \frac{N}{R} \rfloor$ samples from each cluster. During the process of selecting $\lfloor \frac{N}{R} \rfloor$ samples from R clusters, we encountered two special cases that need additional designs. We list them below.

Case 1: N is no smaller than the size of text-similarity-based selected speech-to-text pairs. We select all text-similarity-based selected speech-to-text pairs and ignore the upper limitation N by skipping MCSS. As a result, all text-similarity-based selected speech-to-text pairs are directly input to CMSN.

Case 2: N is smaller than the size of text-similarity-based selected speech-to-text pairs, and there exists a cluster with a size smaller than $\lfloor \frac{N}{R} \rfloor$. We address this case by a greedy-based sample selection algorithm. It greedily selects all samples in a cluster if the cluster size is smaller than a minimum requirement, which is initialized as $r_{min} = \lfloor \frac{N}{R} \rfloor$ and r_{min} is then updated. Finally, the remaining clusters with cluster sizes that are greater than r_{min} will select r_{min} samples from each remaining cluster. The algorithm is detailed in Algo. 1.

Algorithm 1 Greedy-Based Sample Selection

Input: R clusters with cluster sizes that are $[l_1, l_2, ..., l_R]$ respectively, and a pre-set expected sampling size N that is smaller than the sum of $[l_1, l_2, ..., l_R]$.

Initialize the number of remaining clusters to be selected, $\hat{R} = R$
Initialize the number of remaining samples to be selected: $\hat{N} = N$
Initialize the minimum size requirement for each cluster: $r_{min} = \lfloor \frac{\hat{N}}{\hat{R}} \rfloor$
Sort $l = [l_1, l_2, ..., l_R]$ from small to large, and represent their sorted index list as $\hat{l}$, where $l[\hat{l}[i]] \leq l[\hat{l}[i+1]]$
Initialize an empty list p to save the cluster index with cluster size smaller than r_{min}
Initialize an empty list r_{sel} to save the selected samples
Initialize $i = 0$
while $l[\hat{l}[i]] < r_{min}$ & $i \neq R$ **do**
    $\hat{l}[i] \to p$
    all samples in $\hat{l}[i]$-th cluster$\to r_{sel}$
    $\hat{N} = \hat{N} - l[\hat{l}[i]]$
    $\hat{R} = \hat{R} - 1$
    $r_{min} = \lfloor \frac{\hat{N}}{\hat{R}} \rfloor$ ▷ Update r_{min}
    $i = i + 1$
end while
Initialize $j = 0$
while $j \neq R$ **do**
    if $\hat{l}[j]$ not in p **then**
        r_{min} samples in $\hat{l}[i]$-th cluster$\to r_{sel}$
        $j = j + 1$
    end if
end while

Output: r_{sel}

A.2 Data Splits and Examples

As for the MiniPS2SLURP dataset construction, we sample 40.5% of SLURP training set for $D^{A\to L,t}$ to train $\Theta^{A\to L,t}$. For $D^{A\to T,t}$ and $D^{A\to T,o}$ used in training $\tilde{\Theta}^{A\to L}$, we use the same 40.5% of the SLURP training set (having totally same speeches to $D^{A\to L,t}$, but no semantics) and full Mini-PS (32255 pairs) respectively to simulate a real collected speech-to-text pair set $D^{A\to T}$.

As for the VoxPopuli2SLUE dataset construction, we sample 45% of SLUE-VoxPopuli fine-tune set for $D^{A\to L,t}$ to train $\Theta^{A\to L,t}$. For $D^{A\to T,t}$ and $D^{A\to T,o}$ used in training $\tilde{\Theta}^{A\to L}$, we use the same 45% of SLUE-VoxPopuli fine-tune set (having totally same speeches to $D^{A\to L,t}$, but no semantics) and full VoxPopuli (182466 pairs) respectively to simulate a real collected speech-to-text pair set $D^{A\to T}$.

We list data examples in Tab. 5.

A.3 License

Our datasets are built on the SLUE-VoxPopuli (Shon et al., 2022) (using CC0 license), VoxPopuli (Wang et al., 2021) (using CC BY 4.0 license), SLURP (Bastianelli et al., 2020) (using CC BY 4.0 license), and Mini-PS (Galvez et al., 2021) (using CC-BY-SA and CC-BY

Dataset	Text Example	Speech Example	Label (Semantics) Example
SLURP	event remaining mona Tuesday	a speech respective to the text	{'scenario': 'calendar'| 'action': 'set'| 'entities': [{'type': 'event_name'| 'filler': 'mona'}| {'type': 'date'| 'filler': 'tuesday'}]}
Mini-PS	are there any other comments but you would don't have a any opposition to the language itself it's fine ok ok any other comments ok should we go	a speech respective to the text	N/A
SLUE-VoxPopuli	better enforcement of the eu animal welfare legislation is one of the key priorities for animal welfare and the commission has invested substantial resources in pursuit of this aim.	a speech respective to the text	Semantics: {'entities': [{'type': ' CARDINAL '| 'filler': 'one'}| {'type': GPE'| 'filler': 'eu'}]}
VoxPopuli	eu pharmaceutical legislation contains a number of tools to facilitate early access to medicines for patients with unmet medical needs.	a speech respective to the text	N/A

Table 5: Sample examples from each data set used in our experiments.

4.0 licenses). Considering these licenses, our usage of these existing datasets is consistent with their licenses. According to these licenses, VoxPopuli2SLUE is CC BY 4.0 license, and MiniPS2SLURP is CC-BY-SA and CC-BY 4.0 licenses.

For the MiniPS dataset, we will release the data once our paper is published, which is allowed by its license.

A.4 Implementation Details

Our work is implemented on SpeechBrain (Ravanelli et al., 2021). The NLU model $\Theta^{T\to L,t}$ is trained by 80% of $D^{T\to L,t}$ and validated by 10% of $D^{T\to L,t}$. The SLU model training also uses the same dataset split ratio. We train NLU for 20 epochs and SLU for 35 epochs, and the parameters performing the best on the validation set will be kept. We set the K-Means cluster numbers as 100 in our both two dataset text embedding spaces, where these text clusters will be used for the MCSS as the text modal cluster results of $D^{T\to L,t}$. For MCSS, we set the numbers of audio clusters, semantic types, and multi-view cluster numbers R as 100, 53, 30 in the MiniPS2SLURP setting and 100, 18, and 30 in the VoxPopuli2SLUE, respectively. Each of the SLU models and NLU models in our experiments consists of an encoder and a decoder. Each SLU encoder is the HuBERT encoder (Hsu et al., 2021). Each NLU encoder is either LSTM (Hochreiter and Schmidhuber, 1997) or BERT (Devlin et al., 2018) encoder. For the SLU and NLU decoders, they are both attentional RNN decoders (Bahdanau et al., 2014). To reproduce our main results for both GloVe-based and SentBERT-based in Tab. 2, we set $\beta = \gamma = \alpha = 0.1$, $\tau = 0.55$, $w^T = w^L = 10$, $w^A = 1$ and $N = 35000$ on MiniPS2SLURP; on VoxPopuli2SLUE, we set $\beta = \gamma = \alpha = 0.1$, $\tau = 0.75$, $w^T = w^L = w^A = 1$ and $N = 5556$. The reported results are obtained from a single run.

A.5 Case Study

We also show case studies of our $\tilde{\Theta}^{A\to L}$ on the two datasets, shown in Table 6.

A.6 Parameter analyses

The parameter analysis of MCSS and CMSN are respectively shown in Figure 6 and Figure 7.

For MCSS, from the Figure. 6, which shows the parameters of the coefficients of MCSS, w^T, w^L and w^A, we can find below.
1. w^T, w^L, and w^A all impact the performance of MCSS. The figure shows performance variant to different weights of w^T, w^L, and w^A.
2. Considering all three views leads to better performance. Among the cases shown in the (2) subfigure, we see that $w^T = w^A = w^L = 1$ leads to better performance than other single-view cases. This shows the benefit of comprehensively considering three views.

For CMSN, we change one parameter at once and keep the rest parameters fixed; we show each of the four parameters on VoxPopuli2SLUE, from

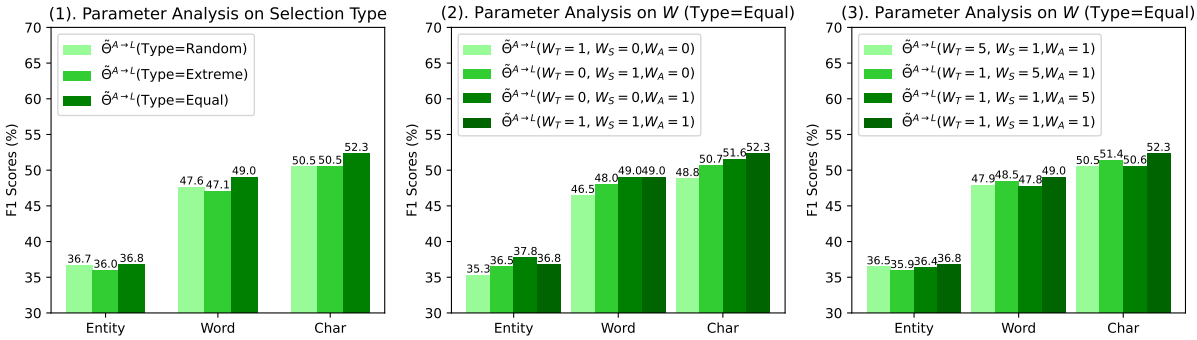

Figure 6: Parameter analysis of MCSS on VoxPopuli2SLUE, where BERT-based $\Theta^{T\to L,t}$ is used. All groups have $\|D^{A\to T,t}\| = 59$ and $\|D^{A\to T,o}\| = 5461$ for fair comparison.

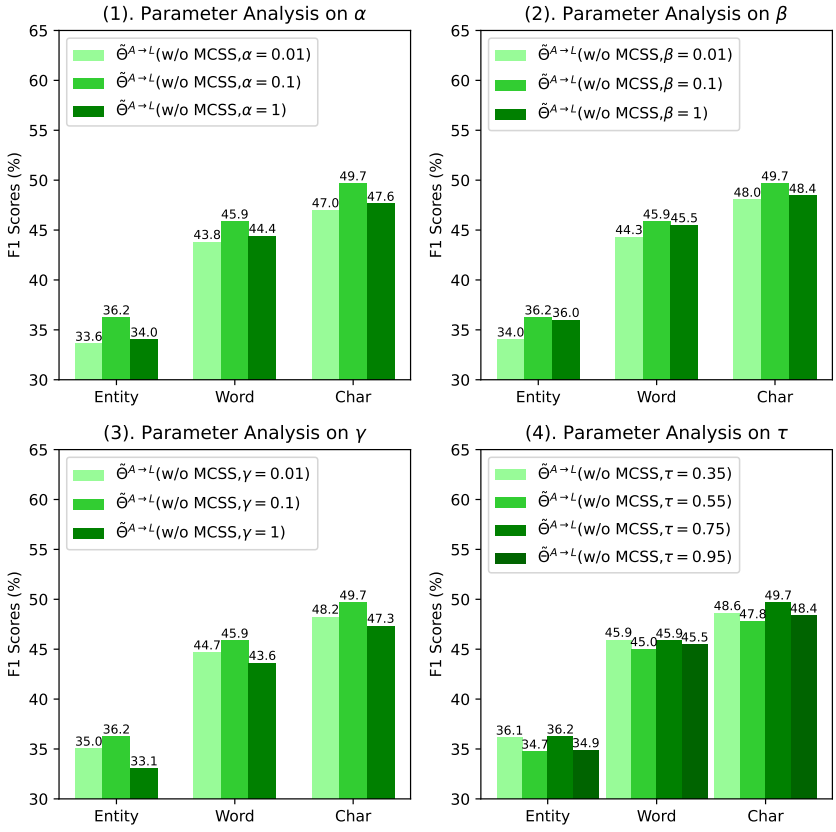

Figure 7: Parameter analysis of CMSN on VoxPopuli2SLUE, where LSTM-based $\Theta^{T\to L,t}$ is used. All groups have $\|D^{A\to T,t}\| = 68$ and $\|D^{A\to T,o}\| = 5489$ for fair comparison.

Audio (Shown by its respective text)	**Ground-Truth Semantic Label**	$\tilde{\Theta}^{A\to L}$ **(w/o CMSN, w/o MCSS) Predicted Label**	$\tilde{\Theta}^{A\to L}$ **(w/o MCSS) Predicted Label**	$\tilde{\Theta}^{A\to L}$ **Predicted Label**
MiniPS2SLURP				
how long does it take to make vegetable lasagna	'scenario': 'cooking', 'action': 'recipe', 'entities': [{'type': 'food_type', 'filler': 'vegetable lasagna'}]	'scenario': 'news', 'action': 'query', 'entities': [{'type': 'news_topic', 'filler': 'election'}, {'type': 'date', 'filler': 'monday'}]	'scenario': 'recommendation', 'action': 'locations', 'entities': [{'type': 'business_type', 'filler': 'restaurant'}]	scenario': 'cooking', 'action': 'recipe', 'entities': [{'type': 'food_type', 'filler': 'cookies'}]
'remind me the meeting with allen on fifteenth march'	'scenario': 'calendar', 'action': 'set', 'entities': [{'type': 'event_name', 'filler': 'meeting}, {'type': 'person', 'filler': 'allen'}, {'type': 'time', 'filler': 'fifteenth march'}]	'scenario': 'calendar', 'action': 'set', 'entities': [{'type': 'event_name', 'filler': 'meeting'}, {'type': 'relation', 'filler': 'wife'}, {'type': 'date', 'filler': 'march'}]	'scenario': 'calendar', 'action': 'set', 'entities': [{'type': 'event_name', 'filler': 'meeting'}, {'type': 'date', 'filler': 'march fifth'}]	'scenario': 'calendar', 'action': 'set', 'entities': [{'type': 'event_name', 'filler': 'meeting'}, {'type': 'person', 'filler': 'allen'}]
can i please have the weather for tomorrow here in costa mesa	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'date', 'filler': 'tomorrow'}, {'type': 'place_name', 'filler': 'costa mesa'}]	'scenario': 'calendar', 'action': 'query', 'entities': [{'type': 'date', 'filler': 'tomorrow'}, {'type': 'time', 'filler': 'eight am'}, {'type': 'date', 'filler': 'tomorrow'}]	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'date', 'filler': 'tomorrow'}, {'type': 'time', 'filler': 'nine am'}]	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'date', 'filler': 'tomorrow'}]
'should i take my raincoat with me now'	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'weather_descriptor', 'filler': 'raincoat'}]	'scenario': 'play', 'action': 'audiobook', 'entities': [{'type': 'media_type', 'filler': 'audiobook'}]	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'weather_descriptor', 'filler': 'rain'}, {'type': 'date', 'filler': 'today'}]	'scenario': 'weather', 'action': 'query', 'entities': [{'type': 'weather_descriptor', 'filler': 'raining'}]
VoxPopuli2SLUE				
second i do not believe in the minsk group but i believe that the eu in the person of the high representative has the capacity to broker the negotiations.	'entities': [{'type': 'gpe', 'filler': 'eu'}, {'type': 'org', 'filler': 'minsk group'}, {'type': 'ordinal', 'filler': 'second'}]	'entities': [{'type': 'gpe', 'filler': 'eu'}, {'type': 'ordinal', 'filler': 'secondly'}, {'type': 'ordinal', 'filler': 'secondly'}]	'entities': [{'type': 'gpe', 'filler': 'eu'}, {'type': 'ordinal', 'filler': 'secondly'}]	'entities': [{'type': 'gpe', 'filler': 'eu'}, {'type': 'ordinal', 'filler': 'second'}]
what can be done to ensure that the revision process goes smoothly and is finalised before one may two thousand and fifteen as specified in article nineteen of the multiannual financial framework regulation so as to avoid losi ng uncommitted amounts from?	'entities': [{'type': 'law', 'filler': 'article nineteen of the multiannual financial framework'}, {'type': 'date', 'filler': 'one may two thousand and fifteen'}]	'entities': [{'type': 'date', 'filler': 'two thousand and twenty'}, {'type': 'date', 'filler': 'two thousand and twenty'}]	'entities': [{'type': 'date', 'filler': 'two thousand and fifty'}	'entities': [{'type': 'date', 'filler': 'two thousand and fifteen'}]

Table 6: Case studies of $\tilde{\Theta}^{A\to L}$ on two datasets, where red fonts show wrong predicted tokens.

which, we find that $\beta = \gamma = \alpha = 0.1$ and $\tau = 0.75$ perform the best.