Interspeech 2022
18-22 September 2022, Incheon, Korea

Contrastive Learning for Improving ASR Robustness in Spoken Language Understanding

Ya-Hsin Chang and Yun-Nung Chen

National Taiwan University, Taipei, Taiwan
r09922066@ntu.edu.tw y.v.chen@ieee.org

Abstract

Spoken language understanding (SLU) is an essential task for machines to understand human speech for better interactions. However, errors from the automatic speech recognizer (ASR) usually hurt the understanding performance. In reality, ASR systems may not be easy to adjust for the target scenarios. Therefore, this paper focuses on learning utterance representations that are robust to ASR errors using a contrastive objective, and further strengthens the generalization ability by combining supervised contrastive learning and self-distillation in model fine-tuning. Experiments on three benchmark datasets demonstrate the effectiveness of our proposed approach.[1]

Index Terms: spoken language understanding, contrastive learning, self-distillation, robustness.

1. Introduction

Intelligent agents such as Apple Siri, Amazon Alexa, and Google Assistant are flourishing with recent advances in speech technology. The core element in these agents is spoken language understanding (SLU), which takes human speech input and extracts semantic information for various tasks, such as intent classification and slot filling. Existing SLU solutions can be categorized into two types: 1) pipeline (or cascade) approaches and 2) end-to-end approaches. Pipeline approaches first use automatic speech recognition (ASR) system to transcribe speech into text, followed by a natural language understanding (NLU) component for the target task, while end-to-end approaches [1] apply a single model which can directly process speech signals and handle the understanding task without considering the text.

Pipeline approaches take many benefits from using textual information: it would be easier to utilize additional resources including large-scale datasets and pre-trained models from the NLP community, relieve the burden of ASR development, and prevent privacy issues about using human voices. However, pipeline methods usually suffer from error propagation – when the ASR hypothesis is incorrect, the erroneous text can mislead the NLU model and hurt the performance. Although ASR systems today have already reached a low word error rate (WER) on data in a controlled environment, a real-world environment still leads to unsatisfied performance. Researchers have explored remedies for ASR errors in two ways: either formulating the ASR error correction as a machine translation from erroneous ASR hypothesis to clean text [2, 3, 4] or adapting the model through masked language modeling (MLM) [5, 6, 7]. Most prior approaches required additional *speech-related* input features such as phoneme sequences [3, 4, 5], lattice graph [7, 8, 9], or N-best hypothesis [10, 11, 12]. Such information may not be easily obtained due to the constraint of ASR systems.

With the rise of BERT [13], pre-trained language models (PLMs) have been dominating the field of NLP. While it is straightforward to adopt PLM in pipeline SLU systems, PLMs are often trained on *clean* text corpus and thus not resistant to ASR errors. In order to learn the invariant representations between manual transcript and erroneous hypothesis, this paper proposes to utilize a contrastive objective to adapt PLM to ASR results with *only textual information*.

Contrastive learning aims at pulling together the feature similarity of positive data pairs and pushing away negative data pairs [14]. In computer vision, positive and negative samples are mostly derived from data augmentation, but due to the characteristics of discreteness in texts, the NLP community does not have a common strategy to create multiple views of a sentence to form positive samples. Prior studies investigated different ways to construct positive pairs, such as back-translation [15], sampling from the same article [16], or the pooled representation from different layers of a model [17]. In spoken scenarios, we naturally take a manual transcript and its associated ASR hypothesis as a positive pair and maximize the similarity between their representations, because they come from the same audio signal. Thus, the learned representations can be more error-robust through contrastive learning in *pre-training*.

In addition, considering the heavily distorted data, we propose a supervised contrastive loss together with a self-distillation strategy during *fine-tuning* in order to further strengthen the generalization capability. Supervised contrastive learning is a variant of contrastive learning [18], where the positive/negative samples are data with the same/different labels, so the annotated target data is required. Results on both vision [18] and language [19] showed improvement and robustness to input noises. Self-distillation [20], or self-knowledge distillation, is a special form of knowledge distillation, where the teacher is the student itself. It simply minimizes Kullback-Leibler (KL) divergence with the model's previous prediction. Without additional information from another model, self-distillation can still demonstrate the regularization effect and prevent overconfidence. We are the first to combine these two techniques for improving robustness in *fine-tuning*.

The contributions of this work are four-fold:

- We propose a novel contrastive objective for pre-training models robust to ASR errors. To our knowledge, we are the first to adopt such modeling techniques to improve robustness with only textual information.
- We propose a novel fine-tuning framework combining supervised contrastive learning and self-distillation.
- The proposed method is flexible and can easily incorporate additional information such as phoneme or lattice.
- Experiments on multiple benchmark datasets demonstrate that the proposed approach is capable of handling noisy text inputs and achieves significant improvement

[1]The codes are released in this github repository https://github.com/MiuLab/SpokenCSE.

10.21437/Interspeech.2022-781

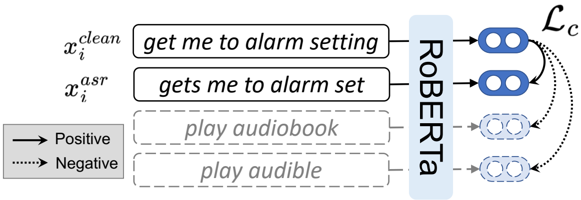

Figure 1: *Pre-training: contrastive learning with the paired ASR noisy transcripts. A positive pair consists of clean data and ASR result from the same audio.*

compared to other methods.

2. Methodology

Our proposed method consists of three elements: (1) A self-supervised contrastive objective for pre-training, (2) supervised contrastive learning, and (3) self-distillation in fine-tuning.

2.1. Self-supvervised contrastive learning

Contrastive learning aims at helping our model distinguish the features invariant to input transformations, including data augmentations and corruptions. To handle ASR errors, we propose to adopt contrastive learning for learning sentence representations invariant to misrecognition. As shown in Figure 1, a pre-trained RoBERTa [21] is continually trained on spoken language corpus by utilizing the paired clean and noisy sentences.

Given a mini-batch of input data of N pairs of texts $B = \{(x_i^{\text{clean}}, x_i^{\text{asr}})\}_{i=1..N}$ representing the clean manual transcript and ASR hypothesis. We first apply the pre-trained BERT and take the last layer of [CLS] to obtain the representation for each sentence as $h = \text{BERT}(x)$. Then we further adjust the sentence representations by the proposed self-supervised contrastive loss [14, 22]:

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

where $\mathbb{P}$ is composed of $2N$ positive pairs of either $(h_i^{\text{clean}}, h_i^{\text{asr}})$ or $(h_i^{\text{asr}}, h_i^{\text{clean}})$, and $s(\cdot,\cdot)$ is a cosine similarity function. The process is illustrated in Figure 1.

Among two terms in the second line of (1), the first term improves the *alignment* between positive pairs with robustness to noise, and the second term promotes *uniformity* in representation space by pushing away features of unrelated samples [23]. Both are known as good characteristics of representations and improve generalization.

To prevent catastrophic forgetting of the PLM, we keep the MLM objective in the pre-training process. Additionally, the prior work revealed the advantage of adaptive pre-training on the same domain of the downstream data [24], so the final proposed pre-training loss $\mathcal{L}_{pt}$ is the weighted sum of a contrastive loss $\mathcal{L}_c$ and an MLM loss $\mathcal{L}_{mlm}$:

$$\mathcal{L}_{pt} = \mathcal{L}_c + \lambda_{mlm}\cdot\mathcal{L}_{mlm}, \tag{2}$$

where λ_{mlm} is a weight to maintain the model's ability of predicting the masked tokens.

2.2. Supervised contrastive learning

Supervised contrastive learning at fine-tuning takes data of the same label as positive samples and pulls their embeddings closer together [18]. In the end, the representations from the same label form a clustering effect and discriminate over different labels by creating margins between them. This objective is similar to the widely-used triplet loss [25], but it can generalize to more than one positive and negative sample and is empirically shown to improve performance and resistance to input noises [19]. We propose to adopt a supervised contrastive loss $\mathcal{L}_{hard}$ to allow the learned representations aligned with their hard labels as illustrated in Figure 2:

$$\mathcal{L}_{hard} = -\frac{1}{N}\cdot\sum_{i}^{N}\sum_{j\neq i}^{N} 1_{y_i=y_j}\log\frac{e^{s(h_i,h_j)/\tau_{sc}}}{\sum_{k\neq i}^{N} e^{s(h_i,h_k)/\tau_{sc}}}. \tag{3}$$

2.3. Self-distillation

Due to ASR errors, some input sentences may no longer retain the semantics of their labels, or shift to some fluent context of another class. For example, when *on* is transcribed into *off* in an IoT control command, the intent can be the exact opposite. Therefore, in order to reduce the impact of label noises in the training set, we propose a self-distillation method.

Self-distillation minimizes KL divergence between the current prediction and the previous one [26, 27], which regularizes the model and eliminates label noise at the same time. We denote $p_i^t = P(y_i \mid x_i, t)$ as the probability distribution of data x_i predicted by the model at the t-th epoch, and its loss function is formulated as:

$$\mathcal{L}_d = \frac{1}{N}\sum_{i}^{N} KL_{\tau_d}(p_i^{t-1}\|p_i^t), \tag{4}$$

and p_i^0 is a one-hot vector of the label y_i. This procedure is illustrated in the right part of Figure 2.

2.4. Self-distilled soft contrastive learning

To relieve the effect of noisy labels in supervised contrastive learning, we add a supplement loss similar to (3) by contrasting the soft label calculated from the previous prediction:

$$\mathcal{L}_{soft} = -\frac{1}{N}\sum_{i}^{N}\sum_{j\neq i}^{N}(p_i^{t-1}\cdot p_j^{t-1})\log\frac{e^{s(h_i,h_j)/\tau_{sc}}}{\sum_{k\neq i}^{N} e^{s(h_i,h_k)/\tau_{sc}}}, \tag{5}$$

This soft target strategy is also investigated in recent self-supervised contrastive learning studies such as ReSSL [28] and SCE [29], where the soft target comes from the similarity between their samples. In the framework shown in Figure 2, our finalized fine-tuning loss $\mathcal{L}_{ft}$ composes of four parts: 1) a cross entropy loss in the original fine-tuning stage $\mathcal{L}_{ce}$, 2) two contrastive learning losses (hard $\mathcal{L}_{hard}$ and soft $\mathcal{L}_{soft}$) and 3) a self-distillation loss $\mathcal{L}_d$ shown as below:

$$\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d\mathcal{L}_d + \lambda_{sc}\big(\mathcal{L}_{hard} + \lambda_d\mathcal{L}_{soft}\big). \tag{6}$$

3. Experiments

3.1. Datasets

Three benchmark datasets are used for evaluating our model: SLURP [30], and two synthesized datasets ATIS and TREC6 from Phoneme-BERT [5]. The statistics are shown in Table 1.

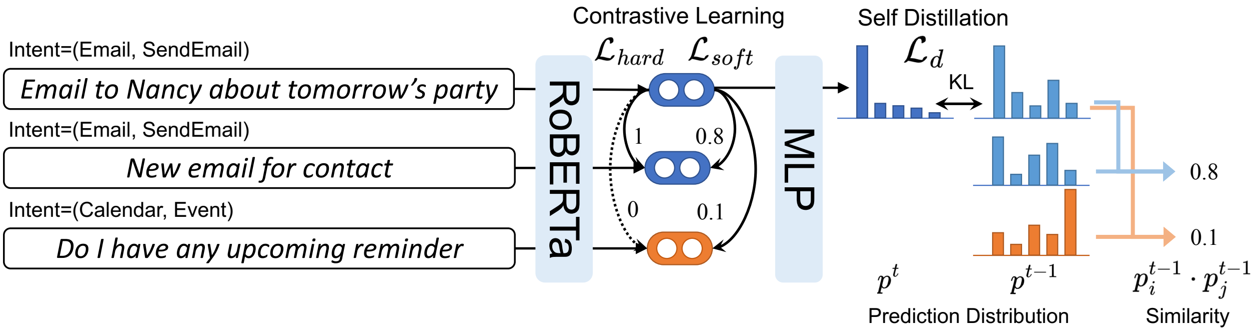

Figure 2: *Fine-tuning: supervised contrastive learning with self-distillation*

Table 1: *Dataset statistics. SLURP test is sub-sampled.*

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 2: *Label noises in SLURP (mislabeled and ambiguous). Model prediction is generated by a fine-tuned RoBERTa.*

Transcripts	Label	Prediction
hello how is your day	(general, quirky)	(general, greet)
let's play music hits	(play, radio)	(play, music)
what date is it	(calendar, query)	(datetime, query)

SLURP is a challenging SLU dataset with various domains, speakers, and recording settings. This paper only focuses on intent detection, where an intent is a (scenario, action) pair, and there are 18 scenarios and 46 actions in total, and the joint accuracy is used as the evaluation metric (both scenario and action are correct). We use two off-the-shelf ASR systems to obtain ASR hypothesis from the provided audio: Google Web API and wav2vec 2.0 [31].[2] The median word error rate (WER) is 25% by Google and 60% by wav2vec, implying the difficulty of performing SLU tasks using ASR hypothesis due to diverse accented speakers and noisy environments in this dataset. Through manual inspection, we find some noisy or incorrect labels in SLURP shown in Table 2, so we sub-sample a test set to ensure its quality for reliable evaluation.[3] Note that the training set may still contain label noises, so the model's robustness to label noises can be still validated in our experiments.

ATIS and TREC6 are two benchmark datasets for flight reservation and question classification respectively. We use the synthesized text released by Phoneme-BERT [5], where the data is synthesized via a TTS model and later transcribed by ASR. Only a subset of data within a certain WER range is kept, and the reported average WER is 29.11% for ATIS and 32.03% for TREC6. We report accuracy as the evaluation metric.

3.2. Experimental setting

We compare our model with three baselines:

- RoBERTa: a RoBERTa-base model directly fine-tuned on the target training data.
- Phoneme-BERT [5]: a RoBERTa-base model further pre-trained on extra corpus with phoneme information. The phoneme sequences are tokenized by a RoBERTa tokenizer with a new token type embedding trained from scratch. Phoneme sequences are generated from the ASR hypothesis via a python toolkit[4] since we do not have access to the phoneme decoder model.
- SimCSE [22]: a state-of-the-art sentence embedding method using contrastive learning. We create positive pairs from two passes of the same ASR hypothesis for calculating $\mathcal{L}_c$ in (1) so that we can better clarify the improvement from learning through manual transcripts in our proposed method.

[2]We use *facebook/wav2vec2-large-960h* trained on LibriSpeech provided on HuggingFace [32].

[3]We only keep samples where an ensemble of 5 RoBERTa models trained on manual transcripts gives the agreed prediction of their labels.

Table 3: *Results on three datasets. Phoneme-BERT† additionally uses phoneme sequences generated by public tool-kit.*

Model	SLURP	ATIS	TREC6
RoBERTa	83.97	94.53	84.08
Phoneme-BERT†	83.78	94.83	85.96
SimCSE	84.47	94.07	84.92
Proposed (pre-train only)	84.51	95.02	85.20
Proposed (pre-train + fine-tune)	**85.26**	**95.10**	**86.36**

We pre-train the model for 10K steps with a batch size 128 on each task data, and fine-tune the model for 10 epochs with early stopping and a batch size 256. In SLURP, two separate classification heads are trained for scenario and action with shared BERT embeddings. We grid search over hyperparameters by metrics of validation set and find that the model is not sensitive, and the final setting is that the mask ratio of MLM is 0.15, $\tau_c = 0.2$, $\lambda_{mlm} = 1$, $\tau_{sc} = 0.2$, $\lambda_{sc} = 0.1$, $\tau_d = 5$, $\lambda_d = 10$. The reported scores are averaged over 5 runs.

3.3. Evaluation results

The evaluation performance is presented in Table 3, where three baselines only focus on pre-training for representation learning, so we additionally show the results only using our proposed pre-training method. The results demonstrate that the proposed contrastive learning in pre-training is useful for handling ASR noises and achieves better performance in most cases. Phoneme-BERT requires additional speech information from phoneme sequences to boost the performance and thus less effective on SLURP, while our model only takes word-level information. Moreover, the proposed fine-tuning framework with contrastive learning and self-distillation further models the uncertainty and improves the performance. In summary, our proposed method is demonstrated effective for SLU and outperforms all baselines on three benchmark datasets.

[4]https://github.com/bootphon/phonemizer

Table 4: *Result on SLURP. WER intervals are separated by quartiles. The reported metric is joint accuracy. Phoneme-BERT uses input text with additional toolkit-generated phoneme sequences.*

Pre-training	Fine-tuning	SLURP WER Interval (Google)					SLURP WER Interval (wav2vec)				
		clean =0	low (0,0.16]	medium (0.16 0.4]	high >0.4	all	low [0,0.25]	medium (0.25,0.5]	high (0.5,0.83]	severe >0.83	all
RoBERTa	Direct	**95.69**	92.41	85.89	56.71	83.97	92.44	80.49	62.06	34.40	68.05
Phoneme-BERT	Direct	94.97	92.34	85.87	57.20	83.78	91.40	79.71	62.60	**36.56**	68.24
SimCSE	Direct	95.55	93.47	**86.82**	57.59	84.47	92.04	81.75	62.98	34.25	68.48
Proposed	Direct	95.54	**93.86**	86.68	**57.72**	**84.51**	**93.02**	**82.56**	**64.22**	36.02	**69.66**
RoBERTa	Proposed	**96.59**	94.27	86.70	57.24	84.87	93.56	81.59	63.29	34.61	68.98
Phoneme-BERT	Proposed	95.61	93.42	86.87	57.50	84.48	92.12	81.46	61.48	33.64	67.89
SimCSE	Proposed	96.57	**94.54**	87.39	58.01	85.25	92.62	80.86	62.13	33.25	67.94
Proposed	Proposed	96.08	94.41	**87.63**	**58.72**	**85.26**	**93.76**	**83.43**	**65.31**	**35.83**	**70.31**

Table 5: *Ablation study of different losses (%).*

$\mathcal{L}_{pt}$	$\mathcal{L}_{ft}$	SLURP	ATIS	TREC6
full	full	**85.26**	**95.10**	**86.36**
no $\mathcal{L}_{mlm}$	full	84.83	93.75	85.32
no $\mathcal{L}_c$	full	85.15	95.00	85.52
full	no $\mathcal{L}_{hard} + \mathcal{L}_{soft}$	85.14	94.83	86.08
full	no $\mathcal{L}_d + \mathcal{L}_{soft}$	84.77	94.75	85.60
full	no $\mathcal{L}_{soft}$	84.81	94.65	86.20

Table 6: *Results on SLURP with parallel data in different ways.*

Pre-train Method	Fine-tune Data	Accuracy
ASR Correction	ASR	83.96
Proposed	ASR	85.26
Proposed	Manual	84.82
Proposed	Manual + ASR	**85.64**

3.4. Analysis of different noise levels

To better investigate the impact of different noise levels, we separate the test set of SLURP into 4 groups according to their WER and show the performance in Table 4. From the upper part of Table 4, our proposed pre-training procedure consistently outperforms all baselines for most cases except clean and severe noise levels. When the input is clean, treating SLU as normal NLU is better, so the original RoBERTa performs best. SimCSE pre-training achieves similar performance with relatively lower WER, but cannot generalize to noisier inputs when we use wav2vec transcripts. Because our proposed self-supervised contrastive learning method allows the model to learn the invariant features by utilizing the relationship between ASR and manual transcripts, the learned spoken representations can be robust even to much noisy inputs. From the lower part of Table 4, our proposed fine-tuning procedure can further boost the performance regardless of pre-trained methods, showing the effectiveness of our framework combining supervised contrastive learning and self-distillation.

3.5. Ablation study

Because the proposed method is composed of multiple elements, we conduct an ablation study to further investigate the effect of each loss in Table 5. The results show that each proposed method (for pre-training and fine-tuning) contributes to the performance positively. It is obvious that MLM in pre-training and self-distillation in fine-tuning play an important role, because MLM pre-training is crucial to adapting the PLM to noisy inputs, and self-distillation prevents the model from overfitting the mismatched text and label pairs. Moreover, self-distillation also reduces the impact of label noises in SLURP and thus brings more contribution.

3.6. Analysis of exploiting manual transcripts

In our proposed procedure, we pre-train on the paired ASR and manual transcripts via contrastive learning. Such information can be exploited in other ways; for example, we can pre-train a sequence-to-sequence model for correcting ASR hypothesis to its manual transcript, and then the encoder can be used in the fine-tuning stage. The upper part of Table 6 shows that the proposed approach can better utilize the paired signal and achieves better performance. One possible reason is that the limited paired data is insufficient for correcting ASR results, while our proposed contrastive learning does not focus on fine-grained text correction but on representation distance for great effectiveness and better efficiency.

Furthermore, our experiments assume that only ASR results are available during fine-tuning for better practice, because the original audio signal may not be manually transcribed for privacy issues. If both manual and ASR transcripts are available in the downstream task, we can also fine-tune the model on both types of data. The lower part of Table 6 shows that additionally utilizing manual transcripts in fine-tuning is beneficial, but taking ASR transcripts is necessary to align with the target inference scenario. Future work can consider how to deal with the scenarios when only clean sentences are available. Also, even without SLU data, our proposed method can still utilize any speech corpus with off-the-shelf ASR in our pre-training stage. Hence, our future work is to validate if other available large speech corpus can further improve the robustness of ASR through our proposed method.

4. Conclusions

This work introduces a novel contrastive objective for learning ASR-robust representations and utilizes supervised contrastive learning and self-distillation to better handle the uncertainty and prevent overfitting. To our knowledge, we are not only the first to utilize contrastive learning for modeling the invariant features between manual and ASR transcripts but also the first to combine self-distillation with supervised contrastive learning for better handling uncertainty. Experiments on three benchmark datasets demonstrate the effectiveness of the proposed framework, showing the great potential of bridging the gap of understanding performance between clean and noisy inputs.

5. References

[1] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-end neural transformer based spoken language understanding," *arXiv preprint arXiv:2008.10984*, 2020.

[2] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "Asr error correction and domain adaptation using machine translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6344–6348.

[3] H. Wang, S. Dong, Y. Liu, J. Logan, A. K. Agrawal, and Y. Liu, "Asr error correction with augmented transformer for entity retrieval." in *Interspeech*, 2020, pp. 1550–1554.

[4] S. Dutta, S. Jain, A. Maheshwari, G. Ramakrishnan, and P. Jyothi, "Error correction in asr using sequence-to-sequence models," *arXiv preprint arXiv:2202.01157*, 2022.

[5] M. N. Sundararaman, A. Kumar, and J. Vepa, "Phoneme-bert: Joint language modelling of phoneme sequence and asr transcript," *arXiv preprint arXiv:2102.00804*, 2021.

[6] C. Wang, S. Dai, Y. Wang, F. Yang, M. Qiu, K. Chen, W. Zhou, and J. Huang, "Arobert: An asr robust pre-trained language model for spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[7] C.-W. Huang and Y.-N. Chen, "Adapting pretrained transformer to lattices for spoken language understanding," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 845–852.

[8] Y. Zou, H. Sun, and Z. Chen, "Associated lattice-bert for spoken language understanding," in *International Conference on Neural Information Processing*. Springer, 2021, pp. 579–586.

[9] C.-W. Huang and Y.-N. Chen, "Learning spoken language representations with neural lattice language modeling," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3764–3769.

[10] H. Wang, J. Chen, M. Laali, K. Durda, J. King, W. Campbell, and Y. Liu, "Leveraging asr n-best in deep entity retrieval," *Proc. Interspeech 2021*, pp. 261–265, 2021.

[11] L. Zhu, W. Liu, L. Liu, and E. Lin, "Improving asr error correction using n-best hypotheses," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 83–89.

[12] K. Ganesan, P. Bamdev, A. Venugopal, A. Tushar *et al.*, "N-best asr transformer: Enhancing slu performance using multiple asr hypotheses," *arXiv preprint arXiv:2106.06519*, 2021.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie, "Cert: Contrastive self-supervised learning for language understanding," *arXiv preprint arXiv:2005.12766*, 2020.

[16] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "Declutr: Deep contrastive learning for unsupervised textual representations," *arXiv preprint arXiv:2006.03659*, 2020.

[17] T. Kim, K. M. Yoo, and S.-g. Lee, "Self-guided contrastive learning for bert sentence representations," *arXiv preprint arXiv:2106.07345*, 2021.

[18] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[19] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.

[20] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3713–3722.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[23] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.

[24] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[25] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009.

[26] S. Yun, J. Park, K. Lee, and J. Shin, "Regularizing class-wise predictions via self-knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 876–13 885.

[27] H. Mobahi, M. Farajtabar, and P. Bartlett, "Self-distillation amplifies regularization in hilbert space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3351–3361, 2020.

[28] M. Zheng, S. You, F. Wang, C. Qian, C. Zhang, X. Wang, and C. Xu, "Ressl: Relational self-supervised learning with weak augmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[29] J. Denize, J. Rabarisoa, A. Orcesi, R. Hérault, and S. Canu, "Similarity contrastive estimation for self-supervised soft contrastive learning," *arXiv preprint arXiv:2111.14585*, 2021.

[30] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[31] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.