



DRI-GAN: A Novel Dual Real Input GAN with Triplet Loss for Cross-Lingual and Noisy SLU

Ankit Kumar¹, Munir Georges²

¹School of Computer Science & Engineering, Galgotias University, India

²AImotion Bavaria, Technische Hochschule Ingolstadt, Germany

anketvit@gmail.com, munir.georges@thi.de

Abstract

Intent detection is a critical task in building spoken language understanding (SLU) systems. We propose a novel semi-supervised Dual Real Input Generative Adversarial Network (DRI-GAN) with triplet loss to enhance the performance of this task. This method effectively leverages both labeled and unlabeled data to achieve superior representation learning. We extract and fuse text embeddings from three locally deployed pre-trained Large Language Models (LLMs) and adapt these embeddings for training the DRI-GAN with triplet loss. Our experiments demonstrate three key findings: (i) In noisy SLU environments, the proposed method outperforms the state-of-the-art by +1.44%. (ii) In zero-shot cross-lingual scenarios, our approach yields substantial accuracy improvements, achieving an absolute gain of 4.19% on MultiATIS++ dataset and 14.60% on MASSIVE dataset. (iii) it achieves higher accuracy without fine-tuning, significantly reducing computational load.

Index Terms: speech recognition, human-computer interaction, natural language understanding, large language models

1. Introduction

Understanding a conversation is crucial for meeting a user's goal in a spoken dialogue system, where intent detection plays a pivotal role in Spoken language understanding (SLU) [1, 2, 3] by accurately interpreting the user's intentions and guiding the interaction. Existing SLU systems have achieved notable success in recent years [4, 5]. However, most current SLU models rely heavily on large amounts of labeled training data, limiting their scalability to low-resource languages or domains with limited data [5]. Acquiring large-scale dataset is a time consuming and costly process. To address this, zero-shot cross-lingual SLU has gained significant attention [6, 7, 8], as it enables models to generalize across languages with minimal human annotation. Multilingual pre-trained models, such as Multilingual BERT (mBERT) [9], have demonstrated strong performance in zero-shot cross-lingual SLU by leveraging large multilingual corpora. Despite these advancements, SLU still faces two major challenges: the lack of sufficient labeled data for rare or low-resource languages and the presence of ASR errors in spoken language transcriptions, which degrade model performance.

Low-resource languages suffer from a scarcity of annotated training data, making it difficult to construct accurate SLU models [10]. Several studies have explored zero-shot cross-lingual SLU [11, 12, 4, 5], but the absence of sufficient supervision limits their effectiveness. While pre-trained LLMs have significantly improved performance in such scenarios, they remain vulnerable to ASR-generated text, which often contains ASR errors. ASR noise introduces challenges such as error propagation, where incorrect transcriptions negatively impact intent

detection [13]. Moreover, models trained on clean text struggle to maintain performance in noisy settings, as they lack robustness to ASR variations.

Numerous studies have been conducted in the past to bridge the gap between reference transcriptions and ASR hypotheses. One approach involves correcting the ASR hypothesis using machine translation techniques [14, 15], while another focuses on adapting the model through masked language modeling [16, 17]. In addition to these methods, the Kullback-Leibler (KL) divergence was applied in [13] to align the predictive distributions of both reference transcriptions and ASR hypotheses, making them more similar. More recently, a technique leveraging contrastive learning was introduced in [18] to align the implicit features of paired reference transcription and noisy hypothesis from an ASR.

In this work, we propose a novel training method designed to address both low-resource intent detection and ASR robustness within a unified approach. Unlike existing methods that tackle these challenges separately through distinct approaches, our solution introduces a semi-supervised Dual Real Input-GAN (DRI-GAN) training method with contrastive learning. This unified strategy enhances performance across both scenarios. Inspired by semi-supervised GANs [19], our method mitigates data scarcity by leveraging a small amount of labeled data alongside large-scale unlabeled data, making it particularly effective in low-resource settings [20, 21]. The DRI-GAN training method employs adversarial training for two distinct purposes: (i) capturing varying noisy acoustic conditions in noisy SLU, and (ii) learning language-independent representations in zero-shot cross-lingual SLU. In the noisy SLU task, the generator G produces embeddings that remain robust to ASR noise while preserving essential semantic information by incorporating clean data in parallel. For cross-lingual SLU, our method utilizes contrastive learning with triplet loss to align multilingual representations of the same utterance while distinguishing them from negative pairs. This explicit feature alignment across noisy and cross-lingual scenarios enhances both robustness and generalizability in real-world SLU applications.

To further improve efficiency and representation quality, we integrate LLMEmbed, a novel embedding extraction approach inspired by [22]. This method fuses text embeddings extracted at multiple depths from three locally deployed pre-trained language models (PLMs): BERT [9], RoBERTa [23], and LLaMA-2 [24] for noisy SLU task and m-BERT [9], XLM-RoBERTa [25], and BGE-Multilingual-Gemma2 [26] for zero-shot cross-lingual task. The embeddings extracted from these models are fused to create a richer representation, which is then used in training. This fusion not only improves feature quality but also significantly reduces computational overhead compared to end-to-end fine-tuning of large language models,

making our method more scalable and resource-efficient. To demonstrate the effectiveness of the proposed DRI-GAN training framework with triplet loss, we experiment with the SLU benchmark datasets: SLURP [27], MultiATIS++ [28], and Massive [6]. Rest of the paper is organized as follows: section 2 describes the proposed methodology, and section 3 covers the experimental evaluation. Finally, we conclude this work.

2. Proposed Methodology

Figure 1 illustrates the training pipeline of the proposed DRI-GAN with triplet loss. The training process consists of three main steps. First, we extract and fuse embeddings from multiple pre-trained LLMs for the reference transcriptions. Second, we perform the same embedding extraction and fusion process for the ASR hypotheses. Finally, the fused embeddings of both the reference transcriptions and ASR hypotheses are fed into the DRI-GAN training method, where triplet loss is applied. This loss function encourages embeddings from the same audio signal (reference transcription and ASR hypothesis) to be more similar while pushing apart embeddings generated from generator. By enforcing this alignment, the model learns error-robust representations, making it more resilient to ASR noise. In zero-shot cross-lingual experiments, the same process is applied, where reference transcriptions and ASR hypotheses are replaced by language pairs. For clarity, the following explanation focuses solely on noisy SLU scenarios to avoid confusion.

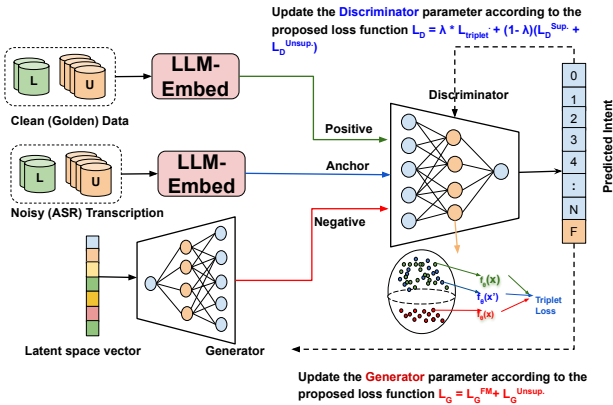


Figure 1: Proposed DRI-GAN method with triplet loss

To extract and fuse the embeddings, we employ LLMEmbed [22], a method that integrates embeddings from multiple pre-trained LLMs at various depths within their networks to boost their robustness and discriminative power. Additionally, since these embeddings are extracted from locally deployed LLMs, our method ensures user data privacy [29] while significantly reducing computational costs [30] making the training process more efficient and scalable.

2.1. Triplet loss for DRI-GAN

Triplet loss, first introduced in [31] and later applied in GAN-based frameworks [32] in 2017, aims to map data points of the same class to be closer in the embedding space while pushing apart data of different classes. A triplet embedding consists of an anchor (x_a), a positive (x_p), and a negative (x_n)

sample. The anchor sample (x_a) represents any class, while the positive sample (x_p) is from the same class as the anchor, and the negative sample (x_n) comes from a different class. In this work, we incorporate triplet loss into our DRI-GAN training in a novel way. As illustrated in Figure 1, we take naturally paired positive samples by considering ASR hypothesis as an anchor $x_i^a (= x_i^{noisy}) = f(x_i^a)$ and reference transcription as a positive $x_i^p (= x_i^{clean}) = f(x_i^p)$ samples. For negative sample, we consider the synthetically generated sample $x_i^n (= G_\theta(z)) = f(x_i^n)$ from the generator. This formulation ensures that the triplet loss explicitly aligns feature representations $f(x)$ of the same utterance, enhancing the model ability to learn noise-robust embeddings. In our experiments, cosine distance was used to compute pairwise distances. The loss term $l_{triplet}^{slu}$ used in training is defined as follows:

$$l_{triplet}^{slu} = [d_{cos}(x_a, x_p) - d_{cos}(x_a, x_n) + m]_+ \quad (1)$$

Here, $d_{cos}(x, y)$ denotes the cosine distance, computed as $d_{cos}(x, y) = 1 - (\frac{x \cdot y}{\|x\| \|y\|})$ and $l_{triplet}^{slu}$ term encouraged to pull the ASR hypothesis close to the reference transcription, which makes GANs more capable of handling variations in data.

2.2. Semi-supervised DRI-GAN training for improved SLU

The proposed DRI-GAN training method builds on SS-GAN [20] by introducing two real inputs into the discriminator. Like SS-GAN, DRI-GAN incorporates generated samples from the generator (G) to the real dataset, and increasing the classifier output dimension from N to $N + 1$ to accommodate the “fake” class. Here, p_r and p_G represent the real and generated data distribution. Discriminator (D) takes both reference and ASR hypothesis as positive and anchor inputs, using triplet loss to explicitly align them while learning to distinguish real samples from generated ones. This dual real input strategy helps to learn better discriminative features and improves its performance in noisy settings. The loss function of D with additional triplet $l_{triplet}^{slu}$ loss can be defined as:

$$L_D = \lambda * l_{triplet}^{slu} + (1 - \lambda) * (L_{D_{sup.}} + L_{D_{unsup.}}) \quad (2)$$

where:

$$L_{D_{sup.}} = \alpha * L_{D_{sup.}}^{noisy} + (1 - \alpha) * L_{D_{sup.}}^{clean} \quad (3)$$

In Eq. 3, $L_{D_{sup.}}^{noisy}$ denotes the supervised loss on ASR hypothesis, and $L_{D_{sup.}}^{clean}$ denotes the supervised loss on reference transcriptions. By setting more value to α , we can give more weightage to noisy SLU task. Reference transcript sample is denoted by x and ASR hypothesis sample is represented as \bar{x} . The supervised loss of D can be calculated as:

$$L_{D_{sup.}}^{noisy} = -\mathbb{E}_{\bar{x}, w \sim p_r} \log[p_m(\hat{w} = w/\bar{x}, w \in (1, \dots, N))] \quad (4)$$

$$L_{D_{sup.}}^{clean} = -\mathbb{E}_{x, w \sim p_r} \log[p_m(\hat{w} = w/x, w \in (1, \dots, N))] \quad (5)$$

By including both terms in the overall training objective, the model can learn to generalize better in real-world conditions where noisy inputs are common.

The unsupervised loss $L_{D_{unsup}}$ for the discriminator is defined as follows:

$$L_{D_{unsup}} = -\mathbb{E}_{\hat{x} \sim p} \log[1 - p_m(\hat{w} = w/\hat{x}, w \in N + 1)] - \mathbb{E}_{\hat{x} \sim G} \log[p_m(\hat{w} = w/\hat{x}, w \in N + 1)] \quad (6)$$

The loss of Generator (G) in DRI-GAN can be calculated as:

$$L_G = L_{G_{ffrm}} + L_{G_{unsup}}. \quad (7)$$

Here, $L_{G_{ffrm}}$ is the fused feature representation matching loss, and $L_{G_{unsup}}$ is the unsupervised loss component. The $L_{G_{ffrm}}$ loss can be defined as:

$$L_{G_{ffrm}} = \beta * L_{G_{ffrm}}^{noisy} + (1 - \beta) * L_{G_{ffrm}}^{clean} \quad (8)$$

$L_{G_{ffrm}}$ encourages the generator to create samples that are statistically similar to real data in the discriminator’s feature space. By setting the value of β , we can adjust the weight to the noisy and clean data. In this paper, we give more weightage to the noisy feature matching as this is the primary task and feature matching on clean transcription helps in the overall training. All hyperparameter settings used in this study are detailed in the Results section.

2.3. Fusing LLM Embeddings with Co-occurrence Pooling

The fusion of LLM embeddings follows the method described in [22]. The term $\mathcal{E}^{(lm, d_{lm})}$ refers to the embedding extracted from the language model $lm \in \{BERT[9], RoBERTa[23], LLaMa2[24]\}$, with d_{lm} representing the depth of the network. For LLaMa2, embeddings are taken from five layers $d_{lm} \in \{1..5\}$, whereas for BERT and RoBERTa, only the last layer ($d_{lm} = 1$) is used. The embedding fusion is performed as follows:

$$\Psi_i = Cat(PN(Cat(EE^T[1 : 7]), \tau), Avg(\mathcal{E}^{(llama2, d_{lm})})) \quad (9)$$

$$E = Cat(\{\widehat{\mathcal{E}}^{(llama2, d_{lm})}\}, \{\mathcal{E}^{(bert, 1)}\}, \{\mathcal{E}^{(roberta, 1)}\}) \quad (10)$$

PN represents a power normalization function applied to balance the power distribution of co-occurrence, where τ is a hyperparameter that adjusts the slope of the PN(.) function.

$$PN(E; \tau) = \tanh(2\tau E) \quad (11)$$

3. Results

3.1. Datasets and Implementation Details

We evaluate the proposed DRI-GAN method for intent detection task on three datasets, SLURP, MASSIVE, and Multi-ATIS++. SLURP [27] is a complex SLU dataset that encompasses a wide range of domains, speakers, and recording environments. The intent classes in SLURP are structured as a pair of (scenario, action), comprising 60 unique combinations derived from 18 scenarios and 46 actions. For cross-lingual evaluation, we employ two benchmark SLU datasets: MASSIVE [6] and MultiATIS++ [28]. MASSIVE is a multilingual text version of SLURP dataset covering 51 languages from 14 language families and in 21 distinct scripts. MultiATIS++ is a multilingual version of the ATIS dataset, supporting nine languages: English (en), Spanish (es), French (fr), German (de),

Table 1: *Dataset statistics.* † denotes the different splits of MultiATIS++ for *hi, tr* language.

Dataset	#intent	Avg. Length	Train	Eval	Test
SLURP	60	6.93	50,628	8690	10,992
MultiATIS++	18	–	4488†	490†	893†
MASSIVE	60	–	11,514	2033	2974

Hindi (hi), Japanese (ja), Portuguese (pt), Turkish (tr), and Chinese (zh). For Hindi and Turkish, the splits for train, dev., and test are 1,440/160/893 and 578/60/715, respectively.

The proposed DRI-GAN method is implemented in PyTorch by extending the original GAN-BERT¹ [20]. For the noisy SLU task, we employed BERT [9], RoBERTa [23], and LLaMa2-7B [24], combining their embeddings into a fused representation of size 4145 using LLMEmbed as described in Section 2.3. In the cross-lingual setting, we utilized m-BERT [9], XLM-RoBERTa [25], and BGE-Multilingual-Gemma2 [26] to create fused embeddings of size 3633. The architectural components of the generator G and discriminator D remain identical to those described in GAN-BERT [20]. Both G and D are implemented as Multi-layer Perceptrons (MLPs) with two hidden layers, activated using leaky ReLU with a dropout rate of 0.2 after each hidden layer. The generator receives a noise vector of size 400, sampled from a normal distribution $\mathcal{N}(0, 1)$, which is passed through the MLP to produce a 4145-dimensional vector, aligning with the dimension of the fused embeddings. These vectors represent the generated (fake) samples. We used hyperparameter values of $\alpha = 0.7$, $\beta = 0.7$, $\lambda = 0.5$, and $\tau = 0.5$. All experiments were conducted using NVIDIA A100 80GB GPUs, with a batch size of 1024. To make results more robust, every experiment is conducted five times with different shuffle and the averaged value is used as final result for comparison.

3.2. Experiments with Noisy-SLU

Table 2: *SLURP English Dataset Evaluation on noisy scenario.*

Model	Noisy _{0.25}	Noisy _{0.60}	Clean
GAN-BERT[20]	82.5	67.10	95.38
Joint-BERT [10]	84.13	70.20	97.12
SpokenCSE [18]	85.26	70.31	95.82
LLMEmbed[22]	80.48	66.83	94.93
DRI-GAN-LLMEmbed	84.88	71.09	96.50
DRI-GAN-LLMEmbed + Triplet Loss	85.71	71.75	96.99

In this section, we evaluate the performance of our proposed method on the noisy SLU task using the SLURP dataset. Table 2 presents a comparative analysis between our model and previous approaches. We employ ASR hypotheses generated by the Google Web API and Wav2Vec 2.0 [33], available through the SpokenCSE² [18]. The median Word Error Rate (WER) is 25% for the Google Web API and 60% for Wav2Vec 2.0. We refer to the ASR hypotheses generated by the ASR engines as Noisy_{wer} and the reference transcripts as Clean. Typically, language models trained exclusively on clean transcripts are sensitive to noisy ASR hypotheses, and training with noisy ASR hypotheses often degrades performance on clean data. However, our model achieves strong performance in both clean and

¹<https://github.com/crux82/ganbert>

²<https://github.com/MiuLab/SpokenCSE>

noisy scenarios, demonstrating its robustness in handling ASR errors without compromising accuracy on clean transcripts. Unlike prior methods that depend on a pre-training and fine-tuning with contrastive learning to boost performance, our approach operates on extracted and fused embeddings without necessitating fine-tuning. This strategy not only reduces computational overhead but also delivers superior results.

3.3. Zero-shot Cross-lingual Experiments

This section covers the experiments in zero-shot cross-lingual settings using MASSIVE, and MultiATIS++ dataset. In the proposed DRI-GAN method, we treat one language as an anchor, the other as positive, and synthetic data as a negative sample for triplet pairing. For cross-lingual transfer learning, we used the weighted objective function as described in Section 2.2. By this, we are essentially telling the generator to prioritize learning and generating embeddings that are closer to higher-weighted language. Even though one language is prioritized, the combined embedding will capture shared intent representations that make sense across related languages. This balance helps in making the model robust for zero-shot performance on other languages. On the other side, the discriminator learns to differentiate between real embeddings and synthetic embeddings generated by the generator. The discriminator will adapt to the blend of features, reinforcing those that are common across both languages. In this way, generator and discriminator interact to create robust feature representations that are not tightly coupled to a specific language, enabling better cross-lingual generalization.

Triplet loss enforces that embeddings of similar intents are closer in the vector space, which further strengthens the multilingual aspect, especially when generalizing to other languages in zero-shot scenario.

Table 3: *Experiment results on MultiATIS++ dataset. Results with ‡ are taken from the corresponding paper.*

Intent Accuracy	pt	zh	ja	hi	tr	AVG
CoSDA [‡] [11]	93.05	78.95	73.25	82.75	80.42	81.68
GL-CLEF [‡] [12]	96.08	87.68	82.84	86.00	83.92	87.30
LAJ-MCL [‡] [4]	97.09	89.03	81.86	84.54	85.45	87.59
FC-MTLF [‡] [5]	97.34	89.53	82.95	86.72	86.02	88.51
Proposed GAN	97.45	93.95	91.93	91.37	88.81	92.70

In Table 3, we present the results of the zero-shot cross-lingual SLU experiment conducted on the MultiATIS++ dataset [28]. For cross-lingual transfer learning, positive pair of two languages is used as pair of (anchor, positive) samples. We extracted the fused embeddings of dimension 3633 as described in Section 2.3, and 3.1. The proposed model was trained on the English-German (en-de) language pair and evaluated on Portuguese (pt), Chinese (zh), Japanese (ja), Hindi (hi), and Turkish (tr) in a zero-shot setting. Our model demonstrates an absolute average improvement of 4.19% over the current state-of-the-art FC-MTLF model [5] across the five target languages.

Figure 2 illustrates the zero-shot performance of our model on the MASSIVE dataset across six Indian languages: Hindi (hi), Telugu (te), Tamil (ta), Bengali (bn), Kannada (kn), and Malayalam (ml). Notably, our approach achieved a maximum absolute improvement of 18.75% in Kannada and a minimum absolute improvement of 8.99% in Hindi compared to the results reported in the MASSIVE paper [6]. Our training method effectively develops robust multilingual intent detection models that generalize to unseen but linguistically or semantically similar languages.

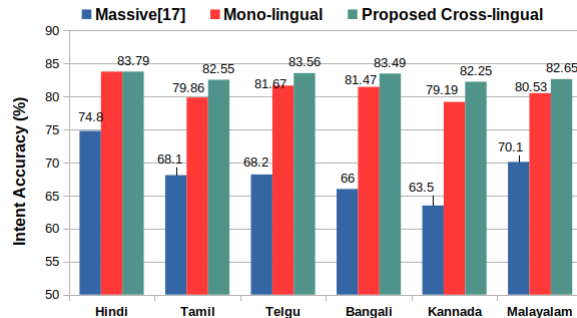


Figure 2: *Zero-shot cross-lingual results on MASSIVE Dataset.*

In Table 4, we present a detailed evaluation of the proposed method across various Indian languages, assessing both mono-lingual and cross-lingual performance. For cross-lingual experiments, we selected Hindi as the source language paired with other target languages. However, in certain instances, such as with the Tamil language, the results did not show improvement over the monolingual method. This suggests that pairing Hindi and Tamil may not be ideal due to their linguistic dissimilarities. In Table 3, the performance differences are subtle, as we are comparing results using the same method, where only slight improvements are expected.

Table 4: *Experiment results on MASSIVE dataset (Indian languages).*

Source/ Target Language		hi	te	ta	bn	kn	ml
Mono-lingual	hi	87.29	79.29	78.65	81.47	76.06	79.35
	te	82.08	84.63	77.98	79.96	79.19	80.20
	ta	83.38	80.63	84.90	79.58	78.28	80.53
	bn	83.79	79.66	78.21	84.26	78.27	78.35
	kn	83.52	81.67	79.86	80.26	84.23	79.42
	ml	81.47	79.89	78.35	77.94	77.94	85.57
Cross-lingual	hi-te	87.42	85.33	80.50	82.62	81.47	82.01
	hi-ta	87.73	82.45	84.70	81.94	80.63	82.25
	hi-bn	87.72	81.81	80.56	85.27	80.33	80.67
	hi-kn	87.66	83.27	82.08	82.58	84.87	81.10
Cross-ling.+ triplet loss	hi-ml	88.00	82.92	81.64	81.98	79.76	85.85
	hi-te	88.03	85.54	81.54	83.09	82.25	82.65
	hi-ta	87.96	83.32	85.03	83.49	81.57	82.62
	hi-bn	87.86	82.68	81.41	85.31	80.60	81.61
	hi-kn	87.70	83.56	82.55	83.49	85.10	82.31
hi-ml	87.90	83.15	81.94	82.65	80.96	85.98	

4. Conclusion

We propose a semi-supervised DRI-GAN method enhanced with triplet loss. Our primary contribution lies in integration of triplet loss within the DRI-GAN architecture, facilitating improved alignment between linguistic and acoustic representations. Our method outperforms state-of-the-art approaches, achieving a maximum improvement of 18.75% on the MASSIVE dataset and 8.98% on the MultiATIS++ dataset in zero-shot cross-lingual scenarios.

5. Acknowledgment

This research was partially supported by the *Exzellenzstiftung Ingolstädter Wissenschaft - Ignaz Kögler*.

6. References

- [1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.
- [2] B. Kim, S. Ryu, and G. G. Lee, “Two-stage multi-intent detection for spoken language understanding,” *Multimedia Tools and Applications*, vol. 76, pp. 11 377–11 390, 2017.
- [3] R. Gangadharaiah, “Joint multiple intent detection and slot labeling for goal-oriented dialog,” 2019.
- [4] S. Liang, L. Shou, J. Pei, M. Gong, W. Zuo, X. Zuo, and D. Jiang, “Label-aware multi-level contrastive learning for cross-lingual spoken language understanding,” *arXiv preprint arXiv:2205.03656*, 2022.
- [5] X. Cheng, W. Xu, Z. Yao, Z. Zhu, Y. Li, H. Li, and Y. Zou, “Fcmf: a fine-and coarse-grained multi-task learning framework for cross-lingual spoken language understanding,” in *Proc. of Inter-speech*, vol. 2, 2023.
- [6] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh *et al.*, “Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages,” *arXiv preprint arXiv:2204.08582*, 2022.
- [7] W. Xu, B. Haider, and S. Mansour, “End-to-end slot alignment and recognition for cross-lingual nlu,” *arXiv preprint arXiv:2004.14353*, 2020.
- [8] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad, “Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark,” *arXiv preprint arXiv:2008.09335*, 2020.
- [9] J. Devlin, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [10] Q. Chen, Z. Zhuo, and W. Wang, “Bert for joint intent classification and slot filling,” *arXiv preprint arXiv:1902.10909*, 2019.
- [11] L. Qin, M. Ni, Y. Zhang, and W. Che, “Cosda-ml: Multilingual code-switching data augmentation for zero-shot cross-lingual nlp,” *arXiv preprint arXiv:2006.06402*, 2020.
- [12] L. Qin, Q. Chen, T. Xie, Q. Li, J.-G. Lou, W. Che, and M.-Y. Kan, “Gl-clef: A global-local contrastive learning framework for cross-lingual spoken language understanding,” *arXiv preprint arXiv:2204.08325*, 2022.
- [13] W. Ruan, Y. Nechaev, L. Chen, C. Su, and I. Kiss, “Towards an asr error robust spoken language understanding system,” 2020.
- [14] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metzger, “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.
- [15] S. Dutta, S. Jain, A. Maheshwari, S. Pal, G. Ramakrishnan, and P. Jyothi, “Error correction in asr using sequence-to-sequence models,” *arXiv preprint arXiv:2202.01157*, 2022.
- [16] 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.
- [17] 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*, vol. 30, pp. 1207–1218, 2022.
- [18] Y.-H. Chang and Y.-N. Chen, “Contrastive learning for improving asr robustness in spoken language understanding,” *arXiv preprint arXiv:2205.00693*, 2022.
- [19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
- [20] D. Croce, G. Castellucci, and R. Basili, “Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples,” in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 2114–2119.
- [21] A. Kumar and M. Georges, “Joint-average mean and variance feature matching (jamvfm) semi-supervised gan with additional-objective training function for intent detection,” in *International Conference on Text, Speech, and Dialogue*. Springer, 2024, pp. 275–287.
- [22] C. ChunLiu, H. Zhang, K. Zhao, X. Ju, and L. Yang, “Llmembed: Rethinking lightweight llm’s genuine function in text classification,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 7994–8004.
- [23] 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.
- [24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [25] A. Conneau, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.
- [26] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation,” 2024.
- [27] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, “Slurp: A spoken language understanding resource package,” *arXiv preprint arXiv:2011.13205*, 2020.
- [28] W. Xu, B. Haider, and S. Mansour, “End-to-end slot alignment and recognition for cross-lingual nlu,” *arXiv preprint arXiv:2004.14353*, 2020.
- [29] Z.-H. Tan, J.-D. Liu, X.-D. Bi, P. Tan, Q.-C. Zheng, H.-T. Liu, Y. Xie, X.-C. Zou, Y. Yu, and Z.-H. Zhou, “Beimingwu: A learnware dock system,” *arXiv preprint arXiv:2401.14427*, 2024.
- [30] M. Xia, T. Gao, Z. Zeng, and D. Chen, “Sheared llama: Accelerating language model pre-training via structured pruning,” *arXiv preprint arXiv:2310.06694*, 2023.
- [31] 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.
- [32] G. Cao, Y. Yang, J. Lei, C. Jin, Y. Liu, and M. Song, “Tripletgan: Training generative model with triplet loss,” *arXiv preprint arXiv:1711.05084*, 2017.
- [33] 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.