Cyclical Contrastive Learning Based on Geodesic for Zero-shot Cross-lingual Spoken Language Understanding

Anonymous ACL submission

Abstract

Owing to the scarcity of labeled training data, Spoken language understanding (SLU) is still a challenging task in low-resource languages. Therefore, zero-shot cross-lingual SLU attracts more and more attention. Contrastive learning is widely applied to explicitly align representations of similar sentences across different languages. However, the vanilla contrastive learning method may face two problems in zero-shot cross-lingual SLU: (1) the consistency between different languages is neglected; (2) each utterance has two different kinds of SLU labels, i.e. slot and intent, the utterances with one different label are also pushed away without any discrimination, which limits the performance. In this paper, we propose Cyclical Contrastive Learning based on Geodesic (CCLG), which introduces cyclical contrastive learning to achieve the consistency between different languages and leverages geodesic to measure the similarity to construct the positive pairs and negative pairs. Experimental results demonstrate that our proposed framework achieves the new state-of-the-art performance on MultiATIS++ and MTOP datasets, and the model analysis further verifies that CCLG can effectively transfer knowledge between different languages[1].

1 Introduction

Spoken Language Understanding (SLU) holds the central position in the task-oriented dialogue systems (Tur and De Mori, 2011; Qin et al., 2019; Xing and Tsang, 2022; Song et al., 2022). The primary objective of SLU is to comprehend and extract relevant information from user utterances. This capability enables the system to discern the user's current objective and generate appropriate responses. SLU comprises two critical sub-tasks: intent detection, which focuses on identifying users' intentions, and slot filling, which entails extracting semantic elements from user queries.

However, the effectiveness of traditional SLU models is intrinsically linked to the availability of extensive annotated data, which poses challenges in scalability. This challenge is particularly evident in the case of low-resource languages, where the lack of substantial labeled datasets exacerbates scalability issues, hindering the seamless deployment and advancement of SLU models. With the demand for language processing solutions extending across diverse linguistic landscapes, the necessity for scalable SLU models that can operate effectively in resource-constrained environments becomes increasingly critical.

To tackle these constraints, the concept of zero-shot cross-lingual SLU generalization has emerged as a central focus of interest and investigation. Recently, mBERT (Devlin et al., 2019) has demonstrated significant advancements in zero-shot cross-lingual SLU. Building upon this work, Liu et al. (2020) introduces an attention-informed mixed-language training approach for cross-lingual SLU. In addition, the exploration of multilingual code-switched settings has been extended by Qin et al. (2020a), which entails aligning a source language with target languages. GL-CLeF (Qin et al., 2022) employs contrastive learning, leveraging bilingual dictionaries to construct multilingual views of the same utterance, then encouraging their representations to be more similar than those negative example pairs. LAJ-MCL (Liang et al., 2022) proposes to model the utterance-slot-word structure using a multi-level contrastive learning framework to facilitate explicit alignment, further enhancing performance. Although existing zero-shot cross-lingual SLU methods have made promising strides by contrastive learning, we identify two main issues:

(1) **The consistency between different languages is neglected.** Although the code-switching method has been applied to construct positive samples in contrastive learning, we find that the consistency between different languages has not been

[1] Our source code and models will be released after review.

Findings of the Association for Computational Linguistics ACL 2024, pages 1806–1816

effectively established. Specifically, the distances between the corresponding samples in different languages are inconsistent, which affects the transfer of knowledge across different languages.

(2) **The utterances with one different label are also pushed away without discrimination.** Traditional contrastive learning methods utilize code-switching to construct the positive samples and negative samples, bringing tokens with the same label and intent label closer together while pushing other the tokens away. However, this can result in a side effect where tokens with only one different label (slot or intent) can be also indiscriminately pushed away, which undoubtedly hampers the representation modeling of contrastive learning, leading to the suboptimal performance.

In this paper, we propose Cyclical Contrastive Learning based on Geodesic (CCLG) to solve these two problems. For the first problem, we introduce two consistency losses, including the cross-lingual consistency loss and the intra-language consistency loss, aiming to boost consistency between different languages. For the second problem, we abandon the conventional approach of directly employing code-switching to construct positive samples and negative samples in contrastive learning. Instead, we utilize geodesic to reconstruct positive and negative samples and employ geodesic-based similarity instead of the traditional similarity metrics, thereby facilitating the learning of representations.

We conduct experiments on MultiATIS++ (Xu et al., 2020) and MTOP (Li et al., 2021), covering nine and six different languages, respectively. The experimental results show that our framework can outperform previous cross-lingual SLU baselines. The model analysis further indicates that our method can transfer knowledge from high-resource languages to low-resource languages. In summary, our work makes three-fold contributions:

- We use cyclical contrastive learning to achieve consistency between different languages.
- We apply geodesic to construct positive and negative samples in contrastive learning, leading to improved representations of tokens.
- Experiment results show that our framework achieves the new state-of-the-art performance on MultiATIS++ and MTOP datasets.

2 Related Works

The related works are introduced from zero-shot cross-lingual SLU and contrastive learning.

2.1 Zero-shot Cross-lingual SLU

Traditional SLU usually focuses on languages with abundant resources, which limits their widespread use. This limitation has sparked growing interest in a novel approach known as zero-shot cross-lingual SLU. The essence of success in this approach lies in tapping into the linguistic insights present in languages with ample resources. By doing so, it opens up exciting possibilities for overcoming challenges posed by limited data in cross-lingual scenarios. Moreover, it extends the reach of SLU to languages that have been previously overlooked, thereby contributing to a more inclusive and adaptable framework in the field of multilingualism.

In recent years, many cross-lingual embeddings, such as mBERT (Devlin et al., 2019), have shown promising results. Liu et al. (2020) propose code-mixing to construct training sentences containing both the source and target phrases, implicitly fine-tuning mBERT. Building upon it, Qin et al. (2020a) proposes multilingual code-switching data augmentation to better align the source language with all target languages. Additionally, van der Goot et al. (2021) suggests three non-English auxiliary tasks to boost cross-lingual transfer. More recently, SoGo (Zhu et al., 2023) highlights the limitations of the conventional code-switching method and proposes a saliency-based substitution approach for extracting keywords as substitutions. In our method, we use cyclical contrastive learning based on geodesic to further transfer the knowledge from the source language to the target language.

2.2 Contrastive Learning

Contrastive learning aims to learn representations of examples via minimizing the distance between positive pairs and maximizing the distance between negative pairs (Saunshi et al., 2019; Chuang et al., 2020; Liu et al., 2022), a concept initially proposed in the field of computer vision (Chopra et al., 2005; Chen et al., 2020; Wang and Liu, 2021). In natural language processing, contrastive learning is utilized for learning the sentence embeddings (Giorgi et al., 2021; Yan et al., 2021), translation tasks (Pan et al., 2021; Ye et al., 2022), and summarization (Wang et al., 2021; Cao and Wang, 2021). Owing to its strong capability in achieving alignment across different languages, contrastive learning has also been used in zero-shot cross-lingual SLU (Liang et al., 2022; Qin et al., 2022). However, we find two main issues with directly utilizing vanilla conservative

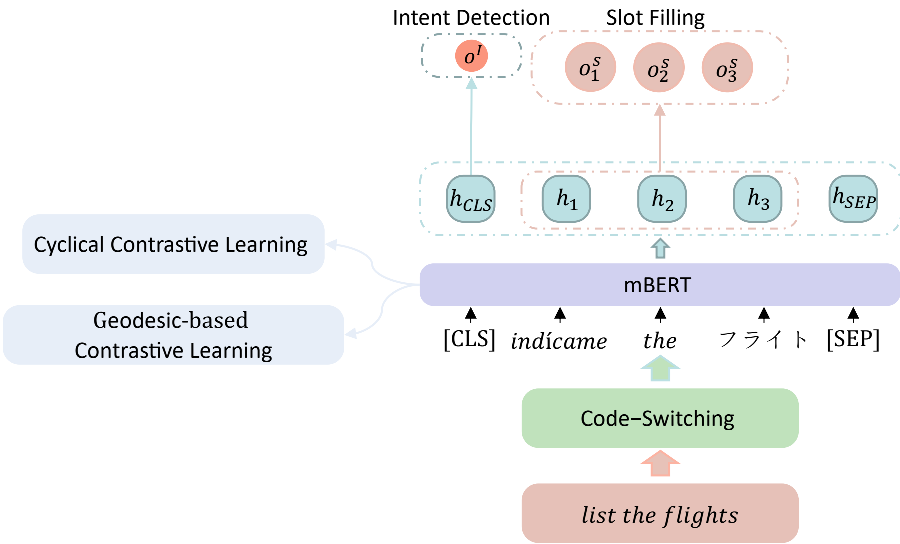

Figure 1: The overview of our approach.

learning in zero-shot cross-lingual SLU. As a result, we propose cyclical contrastive learning based on geodesic to tackle these two issues.

3 Background

SLU comprises two core subtasks, including intent detection and slot filling. Given the input utterance $x = (x_1, x_2, \ldots, x_n)$, where n denotes the length of x, intent detection is treated as a classification task, producing the intent label $\boldsymbol{o}^I$, and slot filling is a sequence labeling task, mapping each utterance x to a slot output sequence $\boldsymbol{o}^S = (o_1^S, o_2^S, \ldots, o_n^S)$. Due to the intrinsic correlation between intent detection and slot filling, it is common to train a unified SLU model capable of jointly handling both tasks, which is formulated as follows:

$$(\boldsymbol{o}^I, \boldsymbol{o}^S) = f(\mathbf{x}) \tag{1}$$

where f denotes the trained model.

Zero-shot cross-lingual SLU task involves training an SLU model on a high-resource source language, such as English, and seamlessly using it on a low-resource target language, such as French. In this scenario, when presented with an instance $\mathbf{x}_{target}$ in the target language, the trained model f can directly generate predictions for both intent and slot values in the target language:

$$\left(\boldsymbol{o}_{target}^I, \boldsymbol{o}_{target}^S\right) = f\left(\mathbf{x}_{target}\right) \tag{2}$$

where $target$ denotes the target language.

4 Method

In this section, we first introduce the Generic SLU Module (Sec. 4.1) and the previous paradigm of utilizing contrastive learning to enhance zero-shot cross-lingual SLU (Sec. 4.2). Then, we introduce the components of our proposed approach, including Cyclical Contrastive Learning (Sec. 4.3) and Geodesic (Sec. 4.4). At last, we introduce the final Training Objective (Sec. 4.5). The overview of our approach is demonstrated in Figure 1.

4.1 Generic SLU Module

Given the input sentence $x = (x_1, x_2, ..., x_n)$, the construction of the input sequence is based on each input utterance by incorporating the specific tokens $\mathbf{x} = (\texttt{[CLS]}, x_1, x_2, ..., x_n, \texttt{[SEP]})$ (Devlin et al., 2019). `[CLS]` serves as the special symbol representing the entire sequence, and `[SEP]` is employed to separate non-consecutive token sequences. Following Qin et al. (2020a), code-switching is applied to leverage the bilingual dictionaries (Lample et al., 2018) in generating multi-lingual code-switched data as input for the model. The representation of the whole utterance, denoted as $\mathbf{H} = (\boldsymbol{h}_{\mathsf{CLS}}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_{\mathsf{SEP}})$, is obtained by utilizing the pre-trained mBERT (Devlin et al., 2019) model.

For the intent detection task, we utilize the utterance representation $\boldsymbol{h}_{\mathsf{CLS}}$ as input to a classification layer in order to derive the predicted intent:

$$\boldsymbol{o}^I = \operatorname{softmax}\left(\boldsymbol{W}^I \boldsymbol{h}_{\mathsf{CLS}} + \boldsymbol{b}^I\right) \tag{3}$$

where $\boldsymbol{W}^I$ and $\boldsymbol{b}^I$ are two trainable matrices.

For the slot filling task, we follow the methods proposed in (Wang et al., 2019; Qin et al., 2022), wherein we use the representation of the first subtoken as the whole word representation and leverage the hidden states to predict each slot:

$$o_t^S = \mathrm{softmax}\left(\boldsymbol{W}^s \boldsymbol{h}_t + \boldsymbol{b}^s\right) \quad (4)$$

where $\boldsymbol{h}_t$ is the representation of the first sub-token of word x_t, $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$ are two trainable matrices.

4.2 Previous Contrastive Paradigm

Contrastive learning has been applied in zero-shot cross-lingual SLU (Qin et al., 2022; Liang et al., 2022). In general, previous methods aim to bring tokens and the corresponding code-switched tokens (positive pairs) closer together while pushing apart tokens and the non-corresponding tokens (negative pairs). And the previous contrastive loss $\mathcal{L}_{\text{CL}}$ can be formulated as follows:

$$\mathcal{L}_{\text{CL}}^{I} = -\sum_{j=1}^{N} \log \frac{s(\boldsymbol{h}_{\text{CLS}}^{j}, \boldsymbol{h}_{\text{CLS}}^{j+})}{\sum_{\boldsymbol{h}_{\text{CLS}}^{j} \neq \boldsymbol{h}_{\text{CLS}}^{j'}}^{B} s(\boldsymbol{h}_{\text{CLS}}^{j}, \boldsymbol{h}_{\text{CLS}}^{j'})} \quad (5)$$

$$\mathcal{L}_{\text{CL}}^{S} = -\frac{1}{n}\sum_{j=1}^{N}\sum_{i=1}^{n} \log \frac{s(\boldsymbol{h}_{i}^{j}, \boldsymbol{h}_{i}^{j+})}{\sum_{\boldsymbol{h}_{i}^{j} \neq \boldsymbol{h}_{i}^{j'}}^{B} s(\boldsymbol{h}_{i}^{j}, \boldsymbol{h}_{i}^{j'})} \quad (6)$$

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_{\text{CL}}^{I} + \mathcal{L}_{\text{CL}}^{S} \quad (7)$$

where $s(\cdot)$ denotes the cosine similarity function, $\boldsymbol{h}_{\text{CLS}}^{+}$ denotes the positive sample of $\boldsymbol{h}_{\text{CLS}}$, $\boldsymbol{h}_{i}^{+}$ denotes the positive sample of $\boldsymbol{h}_i$, B denotes the mini-batch of original and code-switched tokens, and N denotes the total number of utterences.

4.3 Cyclical Contrastive Learning

Inspired by previous work (Goel et al., 2022), to improve the consistency between different languages, we introduce two additional consistency losses, including the cross-lingual consistency loss and the intra-language consistency loss.

The cross-lingual consistency loss $\mathcal{L}_{\text{CCL}}^{C}$ is applied to reduce the discrepancy in similarity scores between the representations of all mismatched pairs of original tokens and code-switched tokens, which can be formulated as follows:

$$\mathcal{L}_{\text{CCL}}^{C} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}(\langle \mathbf{H}_j, \overline{\mathbf{H}}_i\rangle - \langle \mathbf{H}_i, \overline{\mathbf{H}}_j\rangle)^2 \quad (8)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product function, and $\overline{\mathbf{H}}$ denotes the representation of the corresponding code-switched utterance.

The intra-lingual consistency loss $\mathcal{L}_{\text{CCL}}^{I}$ is employed to reduce the discrepancy in the similarity scores between the representations of all the original token pairs and corresponding code-switched token pairs, which can be formulated as follows:

$$\mathcal{L}_{\text{CCL}}^{I} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}(\langle \mathbf{H}_j, \mathbf{H}_i\rangle - \langle \overline{\mathbf{H}}_i, \overline{\mathbf{H}}_j\rangle)^2 \quad (9)$$

The final cyclical contrastive learning loss $\mathcal{L}_{\text{CCL}}$ is the sum of $\mathcal{L}_{\text{CCL}}^{C}$ and $\mathcal{L}_{\text{CCL}}^{I}$:

$$\mathcal{L}_{\text{CCL}} = \mathcal{L}_{\text{CCL}}^{C} + \mathcal{L}_{\text{CCL}}^{I} \quad (10)$$

4.4 Geodesic

In the previous contrastive paradigm, only the tokens with the same two labels, including intent and slot, are regarded as the positive pairs. Therefore, the tokens with only one different label (slot or intent) are also pushed apart without discrimination, which limits the overall performance. To solve this problem, we use geodesic to discriminate positive pairs in contrastive learning.

The representations of tokens are often embedded within a high-dimensional manifold, and our objective is to gauge the geodesic distance between two points along this manifold. However, calculating the precise geodesic distance proves challenging in the absence of explicit knowledge regarding the manifold's structure (Kimmel and Sethian, 1998). To address this, we resort to leveraging the K-NN graph (Cover and Hart, 1967) as an approximation to the manifold structure (Surazhsky et al., 2005; Chowdhury et al., 2022). Within this graph, each token $\boldsymbol{h}_i$ constitutes a node, and connections are established between nodes such that each node links to at most k other nodes.

Specifically, a directed edge is established from the node $\boldsymbol{h}_i$ to node $\boldsymbol{h}_j$ if $\boldsymbol{h}_j$ is one of the k nearest neighbors of $\boldsymbol{h}_i$. The weight of each edge $d(\boldsymbol{h}_i, \boldsymbol{h}_j)$ is defined utilizing the cosine similarity:

$$d(\boldsymbol{h}_i, \boldsymbol{h}_j) = 1 - \boldsymbol{h}_i \boldsymbol{h}_j^{\top} \quad (11)$$

Finally, we employ the shortest path algorithm Djikstra (Dijkstra, 1959) to compute the length of the shortest path between the two token representations along the obtained weighted directed graph, serving as the final geodesic distance $\mathcal{G}(\boldsymbol{h}_i, \boldsymbol{h}_j)$.

For a token $\boldsymbol{h}_i$, we define the k tokens with the closest geodesic distance from the code-switched tokens as its positive samples P_i:

$$P_i = \left\{\boldsymbol{p}_i^k\right\} = \arg\underset{k}{\mathrm{topk}}\mathcal{G}(\boldsymbol{h}_i, \boldsymbol{h}_j) \quad (12)$$

In vanilla contrastive learning, for negative samples with only one different label and those with two different labels, the push operation for negative samples is indistinguishable, which clearly undermines the model to learn the correct representations. As a result, we use the geodesic distance to differentially push negative samples away. The similarity $S_G(\boldsymbol{h}_i, \boldsymbol{h}_j)$ between different tokens is:

$$S_G(\boldsymbol{h}_i, \boldsymbol{h}_j) = \exp(\boldsymbol{h}_i \boldsymbol{h}_j^\top \cdot \log \frac{1}{\exp(\mathcal{G}(\boldsymbol{h}_i, \boldsymbol{h}_j) + 1)}) \tag{13}$$

By considering the relationships between negative samples while maximizing mutual information, we believe $S_G(\boldsymbol{h}_i, \boldsymbol{h}_j)$ is more beneficial than the conventional similarity function. The geodesic-based contrastive learning loss $\mathcal{L}_{\text{GCL}}$ are as follows:

$$\mathcal{L}^I_{\text{GCL}} = -\sum_{j=1}^{N} \log \frac{\sum\limits_{\boldsymbol{p}^k_{\texttt{CLS}} \in P_{\texttt{CLS}}} \exp(\boldsymbol{h}^j_{\texttt{CLS}}, \boldsymbol{p}^k_{\texttt{CLS}})}{\sum^B_{\boldsymbol{h}^j_{\texttt{CLS}} \neq \boldsymbol{h}^{j'}_{\texttt{CLS}}} S_G(\boldsymbol{h}^j_{\texttt{CLS}}, \boldsymbol{h}^{j'}_{\texttt{CLS}})} \tag{14}$$

$$\mathcal{L}^S_{\text{GCL}} = -\frac{1}{n} \sum_{j=1}^{N} \sum_{i=1}^{n} \log \frac{\sum\limits_{\boldsymbol{p}^k_i \in P_i} \exp(\boldsymbol{h}^j_i, \boldsymbol{p}^k_i)}{\sum^B_{\boldsymbol{h}^j_i \neq \boldsymbol{h}^{j'}_i} S_G(\boldsymbol{h}^j_i, \boldsymbol{h}^{j'}_i)} \tag{15}$$

$$\mathcal{L}_{\text{GCL}} = \mathcal{L}^I_{\text{GCL}} + \mathcal{L}^S_{\text{GCL}} \tag{16}$$

4.5 Trainig Objective

Following previous work (Qin et al., 2020b, 2022), the intent detection objective $\mathcal{L}_{\text{I}}$ and the slot filling objective $\mathcal{L}_{\text{S}}$ are computed as follows:

$$\mathcal{L}_{\text{I}} = -\sum_{i=1}^{n_I} \hat{\mathbf{y}}^I_i \log \left(\mathbf{o}^I_i\right) \tag{17}$$

$$\mathcal{L}_{\text{S}} = -\sum_{j=1}^{n} \sum_{i=1}^{n_S} \hat{\mathbf{y}}^{i,S}_j \log \left(\mathbf{o}^{i,S}_j\right) \tag{18}$$

where $\hat{\mathbf{y}}^I_i$ denotes the gold intent label, $\hat{\mathbf{y}}^{i,S}_j$ denotes the gold slot label for the j-th token, n_I denotes the number of gold intent labels, and n_S denotes the number of gold slot labels.

The final training objective $\mathcal{L}$ is as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{I}} + (1-\alpha)\mathcal{L}_{\text{S}} + \lambda\mathcal{L}_{\text{CCL}} + \gamma\mathcal{L}_{\text{GCL}} \tag{19}$$

5 Experiments

5.1 Datasets and Metrics

We primarily conduct our experiments on two public cross-lingual SLU benchmark datasets, including the MultiATIS++ (Xu et al., 2020) dataset and the MTOP (Li et al., 2021) dataset.

MultiATIS++[2] dataset is the broadened version of the Multilingual ATIS (Upadhyay et al., 2018) dataset, whose statistics are shown in Table 1. This extension includes human-translated data for an additional six languages: Spanish (es), German (de), Chinese (zh), Japanese (ja), Portuguese (pt), and French (fr), complementing the original languages, Hindi (hi) and Turkish (tr). The dataset comprises 4,478 utterances in the training set, 500 in the validation set, and 893 in the test set, with a total of 18 intents and 84 slots for each language.

Language	**Utterances** train	valid	test	**Intent types**	**Slot types**
hi	1440	160	893	17	75
tr	578	60	715	17	71
others	4488	490	893	18	84

Table 1: Statistics of MultiATIS++ dataset.

MTOP[3] is compiled from interactions between humans and assistant systems, with statistics presented in Table 2. MTOP comprises over 100,000 human-translated utterances in six languages (English (en), German (de), Spanish (es), French (fr), Thai (th), Hindi (hi)) across eleven domains. For a fair comparison, we Liang et al. (2022) to use the flat version, divided into 70:10:20 percentage splits for the training set, validation set, and test set.

Number of Total Utterances en	de	fr	es	hi	th	**Intent types**	**Slot types**
22288	18788	16584	15459	16131	15195	117	78

Table 2: Statistics of MTOP dataset.

Consistent with prior research (Qin et al., 2022; Zhu et al., 2023; Cheng et al., 2023), accuracy serves as the metric for evaluating intent detection, and F1 score is applied to assess slot filling performance. Moreover, overall accuracy is utilized for sentence-level semantic frame parsing evaluation.

5.2 Implementation Details

Following Qin et al. (2022), we utilize the base case of the multilingual BERT (mBERT)[4](Devlin et al., 2019), featuring $N = 12$ attention heads and $M = 12$ transformer blocks. The learning rate is set to 5×10^{-7} and the total batch size is set to

[2] https://github.com/amazon-science/multiatis
[3] https://fb.me/mtop_dataset
[4] https://github.com/google-research/bert/blob/master/multilingual.md

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	AVG
ZSJoint‡ (Chen et al., 2019)	98.54	90.48	93.28	94.51	77.15	76.59	94.62	73.29	84.55	87.00
CoSDA† (Qin et al., 2021)	95.74	94.06	92.29	77.04	82.75	73.25	93.05	80.42	78.95	87.32
GL-CLEF* (Qin et al., 2022)	98.77	97.53	97.05	97.72	86.00	82.84	96.08	83.92	87.68	91.95
LAJ-MCL* (Liang et al., 2022)	98.77	98.10	98.10	98.77	84.54	81.86	97.09	85.45	89.03	92.41
DiffSLU* (Mao and Zhang, 2023)	98.86	98.17	98.21	98.93	86.66	82.65	97.21	85.98	89.46	92.90
SoGo* (Zhu et al., 2023)	98.89	98.45	98.15	97.74	83.87	84.75	97.73	85.53	89.10	92.69
FC-MTLF* (Cheng et al., 2023)	98.97	98.21	98.36	99.01	86.72	82.95	97.34	86.02	89.53	93.01
CCLG (ours)	**99.35**	**98.51**	**98.94**	**99.43**	**87.32**	**85.53**	**98.79**	**86.48**	**89.97**	**93.81**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
ZSJoint‡ (Chen et al., 2019)	95.20	74.79	76.52	74.25	52.73	70.10	72.56	29.66	66.91	68.08
CoSDA† (Qin et al., 2021)	92.29	81.37	76.94	79.36	64.06	66.62	75.05	48.77	77.32	73.47
GL-CLEF* (Qin et al., 2022)	95.39	86.30	85.22	84.31	70.34	73.12	81.83	65.85	77.61	80.00
LAJ-MCL* (Liang et al., 2022)	96.02	86.59	83.03	82.11	61.04	68.52	81.49	65.20	82.00	78.23
DiffSLU* (Mao and Zhang, 2023)	96.16	86.72	85.48	84.26	73.04	74.12	82.52	68.14	83.12	81.51
SoGo* (Zhu et al., 2023)	95.42	87.46	87.01	84.45	74.25	76.69	83.91	67.04	78.53	81.64
FC-MTLF* (Cheng et al., 2023)	96.21	86.87	85.66	84.62	73.18	74.24	82.68	68.22	83.16	81.65
CCLG (ours)	**96.83**	**88.01**	**87.45**	**85.22**	**74.97**	**77.19**	**84.17**	**68.98**	**83.82**	**82.96**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
ZSJoint‡ (Chen et al., 2019)	87.23	41.43	44.46	43.67	16.01	33.59	43.90	1.12	30.80	38.02
CoSDA† (Qin et al., 2021)	77.04	57.06	46.62	50.06	26.20	28.89	48.77	15.24	46.36	44.03
GL-CLEF* (Qin et al., 2022)	88.02	66.03	59.53	57.02	34.83	41.42	60.43	28.95	50.62	54.09
LAJ-MCL* (Liang et al., 2022)	89.81	67.75	59.13	57.56	23.29	29.34	61.93	28.95	54.76	52.50
DiffSLU* (Mao and Zhang, 2023)	90.06	68.02	59.84	58.08	35.12	43.06	63.04	29.32	55.08	55.74
SoGo* (Zhu et al., 2023)	90.54	72.26	61.05	57.88	39.90	46.95	64.23	29.14	51.31	57.02
FC-MTLF* (Cheng et al., 2023)	91.58	69.54	61.43	59.62	36.86	44.64	64.55	30.86	56.52	57.29
CCLG (ours)	**91.97**	**74.91**	**62.43**	**59.99**	**40.43**	**47.98**	**64.95**	**31.56**	**57.83**	**59.12**

Table 3: Experiment Results on the MultiATIS++ dataset. We report both individual and average (AVG) results. Results with "*" are obtained from the respective published paper, results with "†" are cited from Qin et al. (2022), and results with "‡" are cited from Liang et al. (2022). The symbol "–" indicates missing results from the published work. Results in **bold** denote our framework significantly outperforms baselines with $p < 0.01$ under t-test.

16. During the training process, the value of label smoothing is set to 0.1, and the dropout rate is set to 0.1. We train the model for 40 epochs, and to avoid overfitting, the training will early-stop if the loss on the development set does not decrease for 10 epochs. We use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.98$, and 4k warm-up updates to optimize parameters. Following the zero-shot setting, we choose the model with the highest overall accuracy based on the English development set and subsequently evaluate on test datasets. For all hyper-parameters, we perform several experiments and select the values with the best performance. α is set to 0.9, λ is set to 0.5, γ is set to 1, and k is set to 5. The experiments are conducted on an NVIDIA A100. Our code is based on PyTorch (Paszke et al., 2019) and Transformers[5](Wolf et al., 2020) framework.

5.3 Baselines

We compare our proposed approach with the following baselines, including ZSJoint (Chen et al., 2019), CoSDA (Qin et al., 2021), GL-CLEF (Qin et al., 2022), LAJ-MCL (Liang et al., 2022), DiffSLU (Mao and Zhang, 2023), SoGo (Zhu et al., 2023), and FC-MTLF (Cheng et al., 2023), whose details are provided in Appendix A.

[5] https://github.com/huggingface/transformers

Methods	Intent Acc	Slot F1	Overall Acc
ZSJoint◇	85.31	67.26	52.15
CoSDA‡	90.72	73.34	58.77
CL-CLEF◇	88.94	79.86	61.24
LAJ-MCL*	91.04	74.50	60.11
CCLG (ours)	**92.42**	**82.24**	**64.36**

Table 4: Average results of all the languages on MTOP. Results with ‡ are cited from Liang et al. (2022), results with * are from the corresponding published paper, results with ◇ are obtained by our re-implementation, and results in **bold** denote our framework significantly outperforms baselines with $p < 0.01$ under t-test.

5.4 Main Results

The results on MultiATIS++ are shown in Table 3 and the results on MTOP are listed in Table 4. From them, we have the following observations:

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	AVG
CCLG (ours)	99.35	98.51	98.94	99.43	87.32	85.53	98.79	86.48	89.97	93.81
w/o Cyclical Contrastive Learning	98.21	97.76	97.11	97.74	86.14	84.15	96.01	84.23	88.13	92.16
w/o Geodesic	98.05	97.23	96.54	97.12	85.22	82.05	95.33	83.24	87.42	91.36
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
CCLG (ours)	96.83	88.01	87.45	85.22	74.97	77.19	84.17	68.98	83.82	82.96
w/o Cyclical Contrastive Learning	96.13	87.11	86.82	84.75	74.23	76.65	83.76	68.33	83.08	82.32
w/o Geodesic	95.13	86.04	85.03	83.76	69.97	72.44	81.03	64.98	77.01	79.49
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
CCLG (ours)	91.97	74.91	62.43	59.99	40.43	47.98	64.95	31.56	57.83	59.12
w/o Cyclical Contrastive Learning	91.13	74.22	62.01	59.56	39.64	47.45	64.33	31.02	56.76	58.46
w/o Geodesic	87.62	65.73	59.14	56.62	34.44	41.02	60.11	28.63	50.14	53.72

Table 5: Ablation study of difference components on the MutliATIS++ dataset.

(1) The methodologies employed in CoSDA, GL-CLEF, LAJ-MCL, and FC-MTLF all incorporate code-switching, and it is evident that they outperform models that do not use this technique, showcasing its effectiveness in enhancing model performance compared to those that do not utilize such strategies. Moreover, our proposed approach goes beyond these established approaches by introducing a novel framework that achieves even greater performance gains. With the relative enhancement of 1.83% in average overall accuracy over the previous state-of-the-art model, our method stands out. This notable improvement can be attributed to our innovative approach based on cyclical contrastive learning based on geodesic.

(2) CCLG obtains notable and consistent advancements across all subtasks, particularly showcasing significant improvements. Its impact is particularly pronounced in low-resource languages compared to high-resource ones. The substantial improvement achieved in these languages surpasses gains observed in other high-resource languages. The success of CCLG in low-resource languages aligns with the original intent of the zero-shot cross-lingual SLU task, which aimed to address challenges in languages with limited training data.

5.5 Ablation Study

To validate the advantages of CCLG from different perspectives, we conduct several ablation studies on the MixATIS++ dataset, the results of which are demonstrated in Table 5.

5.5.1 Effect of Cyclical Contrastive Learning

CCLG makes a pivotal contribution through its innovative cyclical contrastive learning, strategically achieving consistency across different languages.

Methods	Intent Acc	Slot F1	Overall Acc
ChatGPT	73.25	61.57	39.16
Vicuna 1.3 (7B)	72.91	60.40	37.05
LLaMA 2 (7B)	72.86	61.20	37.28
CCLG (ours)	**93.81**	**82.96**	**59.12**

Table 6: Results of LLMs on the MutliATIS++ dataset.

To meticulously evaluate the impact of this module, we conduct an ablation study by excluding $\mathcal{L}_{\text{CCL}}$ in Eq. 19, as denoted by "w/o Cyclical Contrastive Learning" in Table 5. A discernible degradation in performance emerges across all metrics for every language when the cyclical contrastive learning module is omitted. We contend that this observed improvement stems from the module's capability to model the consistency between different languages, particularly beneficial for low-resource languages facing the data scarcity challenges.

5.5.2 Effect of Geodesic

To bolster the effectiveness of geodesic, we conduct an ablation study by excluding $\mathcal{L}_{\text{GCL}}$ in Eq. 19. This configuration is denoted as "w/o Geodesic" in Table 5. Significantly, our findings reveal a decline in performance across all metrics for each language, underscoring the importance of geodesic in constructing positive and negative samples in contrastive learning. This ensures a robust and reliable model performance in real-world applications.

5.6 Comparison with Large Language Models

As demonstrated in Table 6, we utilize the evaluation methodology introduced by He and Garner (2023) to assess the performance of ChatGPT (OpenAI, 2023), Vicuna 1.3 (7B) (Zheng et al., 2023), and LLaMA 2(7B) (Touvron et al., 2023). In this

		show	flights	from	burbank	to	st.	louis	on	monday
	Text (En):	show	flights	from	burbank	to	st.	louis	on	monday
Ref.	Intent:	atis_flight								
	Slot:	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	I-toloc.city_name	O	B-depart_date.day_name
GL-CLeF	Intent:	atis_flight								
	Slot:	O	O	O	B-fromloc.city_name	O	O	O	O	B-depart_date.day_name
FC-MTLF	Intent:	atis_flight								
	Slot:	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	O	O	B-depart_date.day_name
CCLG	Intent:	atis_flight								
	Slot:	O	O	O	B-flight_stop	O	O	B-fromloc.city_name	O	B-toloc.city_name
	Text (De):	Zeige	Flüge	von	Burbank	nach	St.	Louis	für	Montag
Ref.	Intent:	atis_flight								
	Slot:	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	I-toloc.city_name	O	B-depart_date.day_name
GL-CLeF	Intent:	atis_airline								
	Slot:	O	O	O	B-fromloc.city_name	O	O	O	O	O
FC-MTLF	Intent:	atis_airline								
	Slot:	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	O	O	O
CCLG	Intent:	atis_flight								
	Slot:	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	I-toloc.city_name	O	B-depart_date.day_name

Table 7: Case study on MultiATIS++ dataset. Text in red denotes the incorrect predictions.

evaluation, the models are presented with 20 examples each. Despite the impressive performance demonstrated by Large Language Models (LLMs) in few-shot and zero-shot learning tasks, a significant performance gap of approximately 20% persists between these models and CCLG in terms of overall accuracy on the MultiATIS++ dataset. This performance disparity is consistently observed across other datasets as well. The observed performance degradation highlights the persistent challenges that language models encounter in understanding spoken language, despite their advanced few-shot and zero-shot learning capabilities. This underscores the urgent need for dedicated efforts in designing effective zero-shot cross-lingual SLU frameworks. Addressing these challenges is not only crucial but also remains an ongoing and vital task for the NLP community. Further exploration and investigation into innovative approaches are warranted to advance state-of-the-art performance.

5.7 Case Study

To further verify the advancements of our model compared to previous methods in zero-shot cross-lingual SLU, we present a case study across different languages. Specifically, we examine English and German as two representative examples. The results in Table 7 reveal notable distinctions in the performance of GL-CLeF, FC-MTLF, and CCLG.

In the case of English, all these models correctly predict the intent. However, as the linguistic complexity increases in German, errors become more pronounced in GL-CLeF and FC-MTLF, while CCLG maintains correct predictions. It exemplifies the robustness and cross-lingual generalizability of CCLG, outperforming its counterparts in accurately predicting intents across diverse languages, without succumbing to increased linguistic complexity, thereby enhancing overall performance.

In terms of slot filling accuracy, GL-CLeF and FC-MTLF show some errors in English, whereas CCLG maintains accuracy. Moving to German, the errors in GL-CLeF and FC-MTLF become more pronounced, while CCLG continues to maintain a high performance. This observed trend highlights the robust nature of CCLG, showcasing its consistent superiority in accurately predicting slots.

6 Conclusion

In this paper, we propose a novel framework CCLG for zero-shot cross-lingual spoken language understanding (SLU), which utilizes cyclical contrastive learning to achieve consistency across different languages and applies geodesic to construct positive samples and negative samples in contrastive learning. Experiments on the MultiATIS++ dataset and the MTOP dataset show that CCLG outperforms the previous best model and achieves a new state-of-the-art performance. Further analysis also demonstrates that our method can indeed transfer knowledge between different languages effectively.

Limitations

While our approach achieves state-of-the-art performance by modifying the traditional contrastive paradigm, we recognize the potential for further enhancements through the incorporation of external knowledge. Given the recent successes observed with LLMs, we anticipate that harnessing LLMs could yield additional improvements in our model's performance. Exploring the integration of LLMs into our framework represents a promising avenue. We leave this aspect for future work.

Ethics Statement

We conducted all experiments using publicly available datasets that are free from offensive content or information with negative social impact. The main objective of this paper is to enhance the model's capacity for understanding, and our model does not generate any uncontrollable output. Hence, we took measures to ensure that our paper adheres to ethical review guidelines. By prioritizing ethical considerations, our aim is to contribute responsibly to the advancement of NLP technology.

References

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proc. of EMNLP*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

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

Xuxin Cheng, Wanshi Xu, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. Fc-mtlf: a fine-and coarse-grained multi-task learning framework for cross-lingual spoken language understanding. In *Proc. of Interspeech*, volume 2.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proc. of CVPR*.

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. 2022. Unsupervised opinion summarization using approximate geodesics. *arXiv preprint arXiv:2209.07496*.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

EW Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

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

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

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

Ron Kimmel and James A Sethian. 1998. Computing geodesic paths on manifolds. *Proceedings of the national academy of Sciences*, 95(15):8431–8435.

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

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proc. of ICLR*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962.

Shining Liang, Linjun Shou, Jian Pei, Ming Gong, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2022. Label-aware multi-level contrastive learning for cross-lingual spoken language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9903–9918.

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

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8433–8440.

Tianjun Mao and Chenghong Zhang. 2023. DiffSLU: Knowledge Distillation Based Diffusion Model for Cross-Lingual Spoken Language Understanding. In *Proc. INTERSPEECH 2023*, pages 715–719.

OpenAI. 2023. ChatGPT (Mar 14 version) [Large language model].

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proc. of ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2078–2087.

Libo Qin, Qiguang Chen, Tianbao Xie, Qixin Li, Jian-Guang Lou, Wanxiang Che, and Min-Yen Kan. 2022. Gl-clef: A global–local contrastive learning framework for cross-lingual spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2677–2686.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197. IEEE.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020a. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020b. Agif: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proc. of ICML*.

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *Proc. of EMNLP*.

Vitaly Surazhsky, Tatiana Surazhsky, Danil Kirsanov, Steven J Gortler, and Hugues Hoppe. 2005. Fast exact and approximate geodesics on meshes. *ACM transactions on graphics (TOG)*, 24(3):553–560.

Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*.

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

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Proc. of ACL Findings*.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proc. of CVPR*.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proc. of EMNLP*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Bowen Xing and Ivor Tsang. 2022. Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs. In *Proc. of EMNLP*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proc. of ACL*.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proc. of NAACL*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing code-switching for cross-lingual slu: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7849–7856.

A Details of Baselines

Here we provide the details of basslines:

(1) ZSJoint: We have re-implemented the zero-shot joint model (Chen et al., 2019) (referred to as ZSJoint), trained on the English training set and directly applied to the test sets of target languages.

(2) CoSDA: Qin et al. (2021) introduces a dynamic code-switching method involving random multilingual token-level replacement. For a fair comparison, we utilize both English training data and code-switching data for fine-tuning.

(3) GL-CLEF: Qin et al. (2022) proposes a global-local contrastive learning framework for explicit alignment, achieving the different granularity alignments, including sentence-level local intent alignment, token-level local slot alignment, and semantic-level global intent-slot alignment.

(4) LAJ-MCL: Liang et al. (2022) introduces a multi-level contrastive learning framework designed for zero-shot cross-lingual SLU.

(5) DiffSLU: Mao and Zhang (2023) introduces a diffusion model and applies knowledge distillation for zero-shot cross-lingual SLU, achieving mutual guidance between intent and slots.

(6) SoGo: Zhu et al. (2023) proposes a semantics-coherent and grammar-coherent method to enhance code-switching method for zero-shot cross-lingual SLU, effectively boosting the performance.

(7) FC-MTLF: Cheng et al. (2023) introduces a framework for cross-lingual SLU, utilizing code-switching for coarse-grained alignment and machine translation for fine-grained alignment.