Code-Switching Can be Better Aligners: Advancing Cross-Lingual SLU through Representation-Level and Prediction-Level Alignment

Zhihong Zhu, Xuxin Cheng, Zhanpeng Chen
Xianwei Zhuang, Zhiqi Huang, Yuexian Zou*
ADSPLAB, School of ECE, Peking University
{zhihongzhu, chengxx, troychen927, xwzhuang}@stu.pku.edu.cn
{zhiqihuang, zouyx}@pku.edu.cn

Abstract

Zero-shot cross-lingual spoken language understanding (SLU) can promote the globalization application of dialog systems, which has attracted increasing attention. While current code-switching based cross-lingual SLU frameworks have shown promising results, they (i) predominantly utilize contrastive objectives to model hard alignment, which may disrupt the inherent structure within sentences of each language; and (ii) focus optimization objectives solely on the original sentences, neglecting the relation between original sentences and code-switched sentences, which may hinder contextualized embeddings from further alignment.

In this paper, we propose a novel framework dubbed REPE (short for **R**epresentation-L**e**vel and **P**rediction-L**e**vel Alignment), which leverages both code-switched and original sentences to achieve multi-level alignment. Specifically, **REPE** introduces optimal transport to facilitate soft alignment between the representations of code-switched and original sentences, thereby preserving structural integrity as much as possible. Moreover, **REPE** adopts multi-view learning to enforce consistency regularization between the prediction of the two sentences, aligning them into a more refined language-invariant space. Based on this, we further incorporate a self-distillation layer to boost the robustness of **REPE**. Extensive experiments on two benchmarks across ten languages demonstrate the superiority of the proposed **REPE** framework.

1 Introduction

Spoken language understanding (SLU) serves as a fundamental component in dialog systems, which involves two tasks: intent detection to classify the intent of user utterances and slot filling to extract useful semantic concepts (Qin et al., 2021; Zhu et al., 2024; Dong et al., 2023a). Recently, massive efforts based on the joint training paradigm (Xing and Tsang, 2022, 2023; Cheng et al., 2023b; Dong et al., 2023b; Zhuang et al., 2024) have shown superior performance in English. Nonetheless, the dependency on extensive labeled training data constrains their applicability to low-resource languages with little or no training data (Dong et al., 2023c), thus hindering the globalization application of dialog systems. Towards this goal, zero-shot cross-lingual SLU gains increasing attention.

Due to the unavailability of low-resource languages (Upadhyay et al., 2018), code-switching (Qin et al., 2020) has been developed to reduce the dependency on machine translation. Technically, it employs bilingual dictionaries to randomly select some words in the sentence to be replaced by their counterparts in other languages. In line with this, numerous zero-shot cross-lingual SLU methods have been proposed (Qin et al., 2022; Liang et al., 2022; Cheng et al., 2023a), yielding promising results. Among them, Qin et al. (2022) incorporated contrastive learning to achieve fine-grained cross-lingual transfer. Based on this, Liang et al. (2022) further proposed a multi-level contrastive learning framework for explicit alignment of utterance-slot-word structure. Recently, Cheng et al. (2023a) integrated with auxiliary task and curriculum learning, obtaining state-of-the-art (SOTA) performance.

Despite the promising progress, we discover existing methods suffer from two main issues: (i) Existing methods (Liang et al., 2022; Qin et al., 2022) employed token-to-token hard contrastive learning objectives to model explicit alignment, potentially disrupting the inherent structural information of sentences, such as inherent phrases or collocations specific to certain languages. (ii) They primarily focus on optimizing objectives based on original sentences, while the correlation between original sentences and code-switched counterparts is ignored, which may lead to the loss of some interactive information and hinder contextualized

*Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 153–160

embeddings from further alignment.

In this paper, we propose a novel framework dubbed REPE to tackle the above two issues. **For the first issue**, we resort to optimal transport (OT) (Peyré et al., 2019) to adaptively model the alignment between the representations of original sentence and code-switched counterpart. In contrast to token-to-token hard contrastive learning, our REPE adaptively considers contextual representations through the alignment matrix, preserving the syntactic structure as much as possible. **For the second issue**, we construct two views from the multilingual pre-trained model (mPLM): the prediction of original and code-switched sentences. By employing multi-view learning (Li et al., 2018), we seek to establish concordance between these two views by minimizing the Kullback–Leibler (Kullback and Leibler, 1951) (KL) divergence, which encourages similar words across different languages to align into a shared latent space. To improve the robustness of the model and prevent over-confidence, we further introduce a self-distillation layer which minimizes KL divergence between the current prediction and the previous one. Experimental results on two benchmarks across ten languages demonstrate that our proposed REPE significantly outperforms previous methods and achieves new SOTA performance, and further analysis verifies the advantages of our REPE.

2 Method

This section introduces the REPE for zero-shot cross-lingual spoken language understanding (SLU), which comprises representation-level alignment (§2.2), prediction-level alignment (§2.3) and self-distillation (§2.4). Figure 1 shows the overview of the proposed REPE framework.

2.1 Task Description

As previously discussed in §1, SLU in dialog systems contains two subtasks: intent detection and slot filling. Since the two subtasks are highly correlated (Goo et al., 2018), it is common to adopt a joint SLU model that can capture shared knowledge. Formally, given an input sentence $\mathbf{x}$ in a target language, zero-shot cross-lingual SLU means the joint model is trained in a source language dataset, *e.g.*, English, and directly applied to the target language datasets, *e.g.*, Chinese:

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

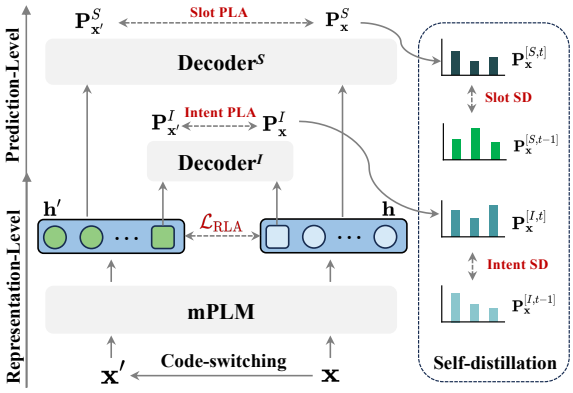

Figure 1: Overview of our proposed REPE.

where $f(\cdot)$ is the joint model; $\mathbf{o}^I$ and $\mathbf{o}^S$ denotes an intent label and a slot sequence. Note that multiple target languages are considered, while only English serves as the source language in our setting.

2.2 Representation-Level Alignment

In existing zero-shot cross-lingual SLU studies, a bunch of works (Liang et al., 2022; Qin et al., 2022) have employed contrastive learning to explicitly align code-switched sentences with original sentences. However, this token-to-token hard alignment disrupts the inherent structure of languages (Zhu et al., 2023). Therefore, we introduce optimal transport (OT) (Peyré et al., 2019) to facilitate soft alignment at the representation level, which aims to find a mapping that transitions probability from one distribution to another with a minimized cost. The OT problem considers two point sets $\mathbf{A} = \{\alpha_i\}_{i=1}^n$ and $\mathbf{B} = \{\beta_i\}_{i=1}^m$, and a transport cost matrix $\mathbf{C}$ with components $\mathbf{C}_{[i,j]} = c(\alpha_i, \beta_j)$ specifying the cost of aligning a pair of points. The goal of OT is to compute a mapping or an alignment matrix $\mathbf{Q}$ that pushes the probability mass of $\mathbf{A}$ toward that of $\mathbf{B}$, while minimizing the sum of costs weighted by the alignments: $\mathcal{L}_{\mathbf{C}} = \sum_{[i,j]} \mathbf{C}_{[i,j]} \mathbf{Q}_{[i,j]}$, where the alignment matrix $\mathbf{Q}$ can be determined using certain OT solution algorithm (*e.g.*, relaxed OT (Kusner et al., 2015), Sinkhorn-Knopp (Sinkhorn and Knopp, 1967) and IPOT (Xie et al., 2020)).

In this work, we denote the original and corresponding code-switched sentence as $\mathbf{x} = \{w_1, w_2, \ldots, w_L\}$ and $\mathbf{x}' = \{w_1, w_2', \ldots, w_L\}$, where w_i' means the replaced source language token by target languages. For a sample $\mathbf{x}$ and its code-switched sentence $\mathbf{x}'$, the multilingual pre-trained language model (mPLM) will produce two different representations $\mathbf{h}, \mathbf{h}'$ (prepended [CLS]

and appended [SEP]). Then, we treat $\mathbf{h}$ and $\mathbf{h}'$ as two point sets and assume each token is uniformly distributed. The cost matrix $\mathbf{C}$ is obtained by computing the cosine distance between contextualized representations in $\mathbf{h}$ and $\mathbf{h}'$. As for the solutions, we use IPOT in this work to obtain the alignment matrix $\mathbf{Q}$, which improves the training speed without degrading the performance as shown in §4.1. The final alignment matrix $\hat{\mathbf{Q}}$ is computed by:

$$\hat{\mathbf{Q}}_{[i,j]} = \text{norm}(\mathbf{Q}_{[i,j]}), \tag{2}$$

where $\text{norm}(\cdot)$ denotes row normalization, which constrains the values to lie between 0 and 1. The value $\hat{\mathbf{Q}}_{[i,j]} = 1$ indicates the extent of alignment between $\mathbf{h}_i$ and $\mathbf{h}'_j$. In this manner, the resulting alignment matrix is used as weak supervision to encourage soft alignment between original and code-switched sentences. The training loss for representation-level alignment is defined as:

$$\mathcal{L}_{\text{RLA}} = -\sum_{[i,j]} \hat{\mathbf{Q}}_{[i,j]} \log(\sigma(1 - \mathbf{C}_{[i,j]})), \tag{3}$$

where σ denotes the sigmoid function, and $1-\mathbf{C}_{[i,j]}$ denotes the consine similarity between $\mathbf{h}_i$ and $\mathbf{h}'_j$.

2.3 Prediction-Level Alignment

For intent detection task, we then feed the whole sentence representations of $\mathbf{h}_{\text{CLS}}$ and $\mathbf{h}'_{\text{CLS}}$ into a classification layer (decoder^I):

$$\mathbf{P}^I_{\mathbf{x}} = \text{softmax}(\mathbf{W}^I \mathbf{h}_{\text{CLS}} + \mathbf{b}^I), \tag{4}$$

$$\mathbf{P}^I_{\mathbf{x}'} = \text{softmax}(\mathbf{W}^I \mathbf{h}'_{\text{CLS}} + \mathbf{b}^I), \tag{5}$$

where $\mathbf{P}^I_{\mathbf{x}}$ and $\mathbf{P}^I_{\mathbf{x}'}$ are intent probability distributions from the original and code-switched sentence, respectively; $\mathbf{W}^I$ and $\mathbf{b}^I$ are intent-specific learnable parameters.

For slot filling task, we similarly feed each hidden state $\mathbf{h}_{[1:-1]}$ and $\mathbf{h}'_{[1:-1]}$ into a classification layer (decoder^S):

$$\mathbf{P}^S_{\mathbf{x}} = \text{softmax}(\mathbf{W}^S \mathbf{h}_{[1:-1]} + \mathbf{b}^S), \tag{6}$$

$$\mathbf{P}^S_{\mathbf{x}'} = \text{softmax}(\mathbf{W}^S \mathbf{h}'_{[1:-1]} + \mathbf{b}^S). \tag{7}$$

The learning objective is to train the classifier to match predicted labels of the original sentence with the ground truth, thus the intent detection loss $\mathcal{L}_I$ and slot filling loss $\mathcal{L}_S$ are defined as:

$$\mathcal{L}_I = \text{CE}(\mathbf{P}^I_{\mathbf{x}}, \mathbf{P}^I), \tag{8}$$

$$\mathcal{L}_S = \frac{1}{L} \sum_{i=1}^{L} \text{CE}(\mathbf{P}^S_{[\mathbf{x},i]}, \mathbf{P}^S_i), \tag{9}$$

where $\text{CE}(\cdot)$ denotes cross-entropy, $\mathbf{P}^I$ and $\mathbf{P}^S_i$ denotes the intent ground truth label and slot ground truth label of i-th token.

On the other hand, we hope the output produced by the decoder^I and decoder^S are language-invariant. Toward this goal, we leverage multi-view learning (Li et al., 2018) to exploit prediction-level alignment from multiple views, which usually contain complementary insights.

Concretely, we consider two distinct views: the probability distribution of original and code-switched sentences. Then, we strive to establish a consensus between these two views, ensuring that the predicted distributions across both two views for each subtask should be as closely aligned as possible:

$$\mathcal{L}_{\text{PLA}} = \underbrace{\text{KL}(\mathbf{P}^I_{\mathbf{x}'} || \mathbf{P}^I_{\mathbf{x}})}_{\text{Intent PLA}} + \underbrace{\text{KL}(\mathbf{P}^S_{\mathbf{x}'} || \mathbf{P}^S_{\mathbf{x}})}_{\text{Slot PLA}}, \tag{10}$$

where $\text{KL}(\cdot)$ denotes Kullback-Leibler divergence (Kullback and Leibler, 1951) to measure the difference between two distributions.

2.4 Self-distillation

To enhance the stability of alignment at both the representation and prediction levels, we introduce a self-distillation (SD) layer to improve the model's robustness. Self-distillation minimizes KL divergence between the current prediction and the previous one (Yun et al., 2020). Specifically, we denote $\mathbf{P}^t_{\mathbf{x}}$ as the probability distribution of the input $\mathbf{x}$ predicted by the model at the t-th epoch, respectively. The whole SD loss $\mathcal{L}_{\text{SD}}$ is combined with its intent- and slot-specific losses expressed as:

$$\mathcal{L}_{\text{SD}} = \underbrace{\text{KL}(\mathbf{P}^{[I,t-1]}_{\mathbf{x}} || \mathbf{P}^{[I,t]}_{\mathbf{x}})}_{\text{Intent SD}} + \underbrace{\frac{1}{L} \sum_{i=1}^{L} \text{KL}(\mathbf{P}^{[S,t-1]}_{[\mathbf{x},i]} || \mathbf{P}^{[S,t]}_{[\mathbf{x},i]})}_{\text{Slot SD}}, \tag{11}$$

where $\mathbf{P}^{[I,t]}_{\mathbf{x}}$ denotes the probability distribution of intent, $\mathbf{P}^{[S,t]}_{[\mathbf{x},i]}$ of slot at i-th token. Note that $\mathbf{P}^{[I,0]}_{x}$ denotes the one-hot vector of the intent label and $\mathbf{P}^{[S,0]}_{[\mathbf{x},i]}$ denotes the one-hot vector of the slot label.

Finally, we train the proposed REPE with a combination of the proposed objectives jointly:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_S + \mathcal{L}_{\text{RLA}} + \mathcal{L}_{\text{PLA}} + \mathcal{L}_{\text{SD}}. \tag{12}$$

3 Experiments

We show the details of the datasets and implementation settings in Appendix §A.1 and §A.2.

Model	MixATIS++			MTOP		
	Intent(Acc)↑	Slot(F1)↑	Overall(Acc)↑	Intent(Acc)↑	Slot(F1)↑	Overall(Acc)↑
CoSDA (Qin et al., 2020)	90.87	68.08	43.15	88.61*	76.85*	58.02*
LAJ-MCL (Liang et al., 2022)	92.41	78.23	52.50	-	-	-
GL-CLeF (Qin et al., 2022)	91.95	80.00	54.09	88.92*	79.84*	61.12*
SoGo_{GL} (Zhu et al., 2023)	92.69	81.64	57.02	-	-	-
FC-MTLF (Cheng et al., 2023a)	93.01	81.65	57.29	-	-	-
REPE (Ours)	**94.17**†	**82.89**†	**58.65**†	**89.46**†	**80.53**†	**63.08**†

Table 1: Main results on MixATIS++ and MTOP. Results with * are from our re-implementation. Results marked with † significantly ($p = 0.05$) improve over all others using the bootstrap confidence interval (Dror et al., 2018).

Model	MixATIS++			MTOP		
	Intent(Acc)↑	Slot(F1)↑	Overall(Acc)↑	Intent(Acc)↑	Slot(F1)↑	Overall(Acc)↑
REPE (Ours)	**94.17**	**82.89**	**58.65**	**89.46**	**80.53**	**63.08**
w/o RLA	88.55↓5.62	80.43↓2.46	52.32↓6.33	83.59↓5.87	77.85↓2.68	56.56↓6.52
w/o PLA	90.28↓3.89	80.86↓2.03	53.36↓5.29	85.08↓4.38	78.28↓2.25	57.21↓5.87
w/o Intent PLA	92.11↓2.06	82.05↓0.84	55.67↓2.98	87.16↓2.30	79.62↓0.91	59.95↓3.13
w/o Slot PLA	92.32↓1.85	81.77↓1.12	56.11↓2.54	87.30↓2.16	79.15↓1.38	60.23↓2.85
w/o SD	92.30↓1.87	81.87↓1.02	56.42↓2.23	87.21↓2.25	79.49↓1.04	60.61↓2.47
w/o Intent SD	93.09↓1.08	82.28↓0.61	57.31↓1.34	88.20↓1.26	79.88↓0.65	61.69↓1.39
w/o Slot SD	93.25↓0.92	82.05↓0.84	57.55↓1.10	88.29↓1.17	79.60↓0.93	61.87↓1.21

Table 2: Ablation study. RLA: representation-level alignment. PLA: prediction-level alignment. SD: self-distillation.

3.1 Main Results

The performance comparison of the proposed REPE framework and baselines are shown in Table 1, from which we have the following observations: **(i)** Our proposed REPE outperforms baselines on both datasets, setting new SOTA in zero-shot cross-lingual SLU tasks, confirming its effectiveness. **(ii)** Statistical tests confirm that REPE's superiority over baselines is significant across evaluation metrics. **(iii)** REPE shows notable gains in accuracy, likely due to soft alignment at the representation level and further refinement at the prediction stage, enhanced by a self-distillation layer that improves cross-lingual transfer. **(iv)** REPE's greater improvement on MixATIS++ is likely because it handles more languages (9 vs. 6) with greater diversity, challenging cross-task transfer. Its success comes from robust multilingual representations and a self-distillation module.

3.2 Ablation Study

We conduct a set of ablation experiments to verify the advantages of our work from different perspectives. From the results in Table 2, we observe that: **(i)** The removal of representation level alignment ("*w/o* RLA") sharply reduces the performance in all evaluation metrics and across both datasets. This indicates that contrasted with hard contrastive learning objectives, employing OT-based soft alignment enhances the quality of representations, which facilitates superior cross-language transfer and preserves the intrinsic structural information within respective languages more effectively. **(ii)** The removal of prediction level alignment ("*w/o* PLA") leads to considerable performance degradation. This implies that performing multi-view learning can facilitate the alignment of predictive information between the original and code-switched sentences, thereby enhancing the complementarity of information. Furthermore, removing either intent PLA or slot PLA ("*w/o* Intent, Slot PLA") results in a decline in overall performance to varying degrees, demonstrating the effectiveness of different submodules. **(iii)** In addition, "*w/o* SD, Intent SD and Slot SD" indicate varying degrees of performance reduction, which proves the effectiveness of self-distillation in our REPE. Given the subjectivity in intent and slot annotation across different languages, our REPE employs self-distillation to mitigate the effects of noisy labels and curb overconfidence, which provides a partial solution.

4 Method Analysis

We further provide insights into the effectiveness of our model by comparing different OT solutions and the potential of leveraging complementary per-

Model	MixATIS++ (Acc)↑	MTOP (Acc)↑	Speed (s)↓
Sinkhorn-Knopp	58.71	63.12	45
Relaxed OT	58.48	62.87	30
REPE (Ours)	58.65	63.08	34

Table 3: Overall accuracy and speed using different OT solutions. Speed: the average training time per epoch.

Model	MixATIS++ (Acc)↑	MTOP (Acc)↑
ORG + CS (Ours)	58.65	63.08
ORG + TRANS	56.12	60.14
ORG + CS + TRANS	60.37	65.09

Table 4: Overall accuracy using different learning views. ORG: original sentence. CS: code-switched sentence. TRANS: translation of original sentence.

spectives for robust cross-lingual representation.

4.1 Impact of OT Solution

In the proposed REPE, we use normalized IPOT to learn the soft alignment between representations of original and code-switched sentences. In this subsection, we compare REPE with other types of OT. From the results in Table 3, we can see Relaxed OT (Kusner et al., 2015) compromises accuracy for increased training speed, whereas the Sinkhorn-Knopp (Sinkhorn and Knopp, 1967) incurs significant training time due to its pursuit of exact solutions. In contrast, the OT solution in our REPE achieves a compromise between the two, enhancing training efficiency while delivering performance comparable to that of the Sinkhorn-Knopp.

4.2 Impact of Learning Views

In this subsection, we add the third view called TRANS to explore the potential of PLA, which is the translation of the original sentence by a machine translation system[1] trained on Europarl[2] corpus. From the results in Table 4, we observe that the translated sentences further enhance the REPE's performance by providing an additional perspective. The translated sentence compensates for the limitations of code-switching, which can occasionally disrupt semantic coherence. Conversely, code-switching introduces more language-independent information compared to the translated sentences. Consequently, the model can learn more robust

[1] https://github.com/facebookresearch/fairseq
[2] https://statmt.org/europarl/

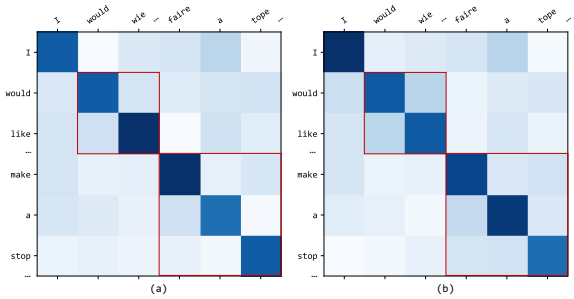

Figure 2: Visualizations of the cosine similarity matrix of the contextualized representations obtained from GL-CLEF and our REPE. (*zoom-in for better view*)

cross-lingual representations by leveraging these complementary perspectives. However, incorporating a complex translation system may be excessive, as large parallel data may not be available for all languages. In a nutshell, our proposed REPE remains straightforward and efficient, which is more suitable for low-resource languages.

4.3 Visualization

To qualitatively demonstrate the superior soft alignment and preservation of syntactic information by the proposed REPE framework, we present an example from the MixATIS++ dataset in Figure 2. It is evident that GL-CLEF achieves commendable representations through contrastive learning for individual tokens, it fails to capture fixed expressions such as "`make a stop`". In contrast, our REPE effectively maintains contextual structural information, successfully recognizing fixed expressions like "`would like`" and "`make a stop`".

5 Conclusion

This work presents REPE, a novel framework for zero-shot cross-lingual SLU. REPE utilizes OT to achieve soft alignment between representations of original and code-switched sentences to preserve structural information within languages. Besides, REPE introduces multi-view learning to predictions of original and code-switched sentences for further alignment and self-distillation to boost the performance. Extensive experiments on two benchmarks show that our REPE outperforms previous models and achieves new SOTA performance.

Limitations

The proposed REPE framework's limitations include the following: (i) The REPE's performance may be affected by the quality of bilingual dictio-

naries used for code-switching. (ii) The effectiveness of the framework is also tied to the quality of the underlying multilingual pre-trained language model, which may not represent all languages equally well. (iii) The soft alignment achieved through optimal transport is an approximation and may not always be perfect. The self-distillation layer, while enhancing robustness, could potentially lead to overfitting if not carefully calibrated.

Ethics Statement

The focus of this article is on a novel framework which leverages both code-switched and original sentences to achieve multi-level alignment, and our model does not have uncontrollable outputs. In addition, all experiments are conducted on publicly available datasets, which do not contain any negative social impact or violations of ethical review.

Acknowledgement

We would like to thank all reviewers for their insightful comments and suggestions to help improve the paper. This paper was partially supported by NSFC (No:62176008).

References

Xuxin Cheng, Wanshi Xu, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023a. FC-MTLF: A Fine- and Coarse-grained Multi-Task Learning Framework for Cross-Lingual Spoken Language Understanding. In *Proc. INTERSPEECH 2023*, pages 690–694.

Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023b. MRRL: Modifying the reference via reinforcement learning for non-autoregressive joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10495–10505, Singapore. Association for Computational Linguistics.

Guanting Dong, Tingfeng Hui, Zhuoma GongQue, Jinxu Zhao, Daichi Guo, Gang Zhao, Keqing He, and Weiran Xu. 2023a. DemoNSF: A multi-task demonstration-based generative framework for noisy slot filling task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10506–10518, Singapore. Association for Computational Linguistics.

Guanting Dong, Zechen Wang, Jinxu Zhao, Gang Zhao, Daichi Guo, Dayuan Fu, Tingfeng Hui, Chen Zeng, Keqing He, Xuefeng Li, Liwen Wang, Xinyue Cui, and Weiran Xu. 2023b. A multi-task semantic decomposition framework with task-specific pre-training for few-shot ner. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 430–440, New York, NY, USA. Association for Computing Machinery.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023c. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*.

Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883.

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.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

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, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. 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-20*, pages 3853–3860. International Joint Conferences on Artificial Intelligence Organization. Main track.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4577–4584. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

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.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR.

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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 159–169.

Bowen Xing and Ivor W Tsang. 2023. Relational temporal graph reasoning for dual-task dialogue language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13876–13885.

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.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1022–1031, New York, NY, USA. Association for Computing Machinery.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19786–19794.

A Dataset and Implementation Details

A.1 Datasets

Following previous works, we conduct experiments on two benchmark datasets: MixATIS++ (Xu et al., 2020) and MTOP (Li et al., 2020). MixATIS++ consists of 9 languages including English (en), Spanish (es), Portuguese (pt), German (de), French (fr), Chinese (zh), Japanese (ja), Hindi (hi), and Turkish (tr). MTOP consists of 6 languages including English (en), German (de), French (fr), Spanish (es), Hindi (hi), and Thailand (th). The statistics of MixATIS++ and MTOP are shown in Table 5 and Table 6, respectively.

Language	Utterances			Intent types	Slot types
	#Train	#Valid	#Test		
hi	1,440	160	893	17	75
tr	578	60	715	17	71
others	4,488	490	893	18	84

Table 5: Statistics of MultiATIS++.

A.2 Implementation Details

Training Settings For a fair comparison, we leverage mBERT (base) (Kenton and Toutanova, 2019) as mPLM (Due to space limitations, results

Utterances (Train&Valid&Test)						Intent	Slot
en	de	fr	es	hi	th	types	types
22,288	18,788	16,584	15,459	16,131	15,195	117	78

Table 6: Statistics of MTOP.

on XLM-R will included in the final version) to encode both original and code-switched sentences. Adam (Kingma and Ba, 2014) is utilized as the optimizer with a learning rate of $3e$-6. When constructing code-switched sentences, bilingual dictionaries of MUSE (Lample et al., 2018)[3] are adopted for code-switching the same as (Qin et al., 2022; Liang et al., 2022) for a fair comparison. Following the zero-shot setting, we use en training set and code-switching set for model training and en validation set for checkpoint saving. We report the average score on the test set of 5 runs with different seeds. We conduct all the experiments on one NVIDIA Tesla P100 GPU.

Evaluation Metrics Following previous works (Qin et al., 2022; Zhu et al., 2023), we evaluate the performance of intent prediction using accuracy (Acc), slot filling using F1 score (F1), and sentence-level semantic frame parsing using overall accuracy (Acc). Higher is better for all metrics.

[3] https://github.com/facebookresearch/MUSE