GL-CLEF: A Global–Local Contrastive Learning Framework for Cross-lingual Spoken Language Understanding

Libo Qin[1]*, Qiguang Chen[1], Tianbao Xie[1], Qixin Li[1], Jian-Guang Lou[2] , Wanxiang Che[1]†, Min-Yen Kan[3]

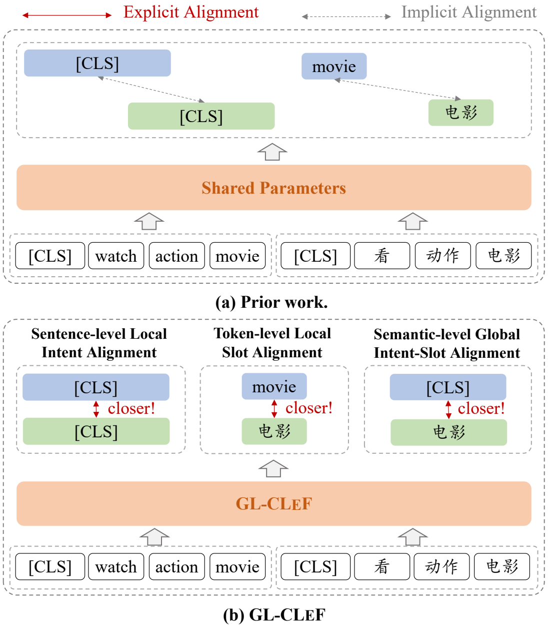
[1]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[2] Microsoft Research Asia, Beijing, China
[3]Department of Computer Science, National University of Singapore
{lbqin,tianbaoxie,qxli,car}@ir.hit.edu.cn; jlou@microsoft.com; kanmy@comp.nus.edu.sg

Abstract

Due to high data demands of current methods, attention to zero-shot cross-lingual spoken language understanding (SLU) has grown, as such approaches greatly reduce human annotation effort. However, existing models solely rely on shared parameters, which can only perform implicit alignment across languages. We present **G**lobal–**L**ocal **C**ontrastive **LE**arning **F**ramework (GL-CLEF) to address this shortcoming. Specifically, we employ contrastive learning, leveraging bilingual dictionaries to construct multilingual views of the same utterance, then encourage their representations to be more similar than negative example pairs, which achieves to explicitly aligned representations of similar sentences across languages. In addition, a key step in GL-CLEF is a proposed `Local` and `Global` component, which achieves a fine-grained cross-lingual transfer (i.e., *sentence-level* `Local` intent transfer, *token-level* `Local` slot transfer, and *semantic-level* `Global` transfer across intent and slot). Experiments on MultiATIS++ show that GL-CLEF achieves the best performance and successfully pulls representations of similar sentences across languages closer.

1 Introduction

Spoken language understanding (SLU) is a critical component in task-oriented dialogue systems (Tur and De Mori, 2011; Qin et al., 2021b). It usually includes two sub-tasks: intent detection to identify users' intents and slot filling to extract semantic constituents from the user's query. With the advent of deep neural network methods, SLU has met with remarkable success. However, existing SLU models rely on large amounts of annotated data, which makes it hard to scale to low-resource languages that lack large amounts of labeled data. To address this shortcoming, zero-shot cross-lingual SLU generalization leverages the labeled training data in high-resource languages to transfer the trained model to a target, low-resource language, which gains increasing attention.



(a) Prior work.

(b) GL-CLEF

Figure 1: (a) Prior work (Implicit Alignment); (b) GL-CLEF (Explicit Alignment). Different color denotes representations across different languages. `[CLS]` represents the sentence representation.

To this end, many works have been explored for zero-shot cross-lingual SLU. Multilingual BERT (mBERT) (Devlin et al., 2019), a cross-lingual contextual pre-trained model from a large amount of multi-lingual corpus multi-lingual corpus, has achieved considerable performance for zero-shot cross-lingual SLU. Liu et al. (2020) further build an attention-informed mixed-language training by generating bi-lingual code-switched data to implicitly align keywords (e.g., slots) between source and target language. Qin et al. (2020) extend the idea to a multilingual code-switched setting, aligning the source language to multiple target languages. This approach currently achieves the

[*] Work done during internship at Microsoft Research Asia and remote visiting at National University of Singapore.

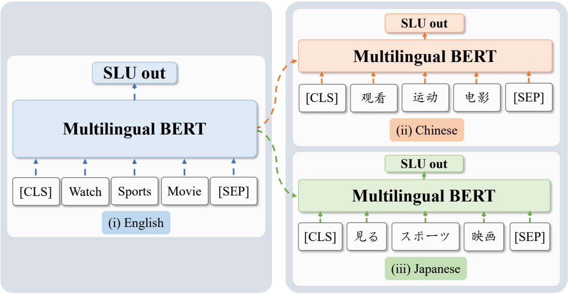
[†] Email corresponding.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2677 - 2686
May 22-27, 2022 ©2022 Association for Computational Linguistics

state-of-the-art performance for zero-shot cross-lingual SLU. Though achieving promising performance, as shown in Figure 1 (a), the above methods solely rely on shared parameters and can only perform implicit alignment across languages, which brings two challenges. First, such implicit alignment process seems to be a black box, which not only seriously affects the alignment representation but also makes it hard to analyze the alignment mechanism. Second, prior work do not distinguish between the varying granularities of the tasks: the intent detection is *sentence-level* and the slot filling is *token-level*, which does not offer fine-grained cross-lingual transfer for *token-level* slot filling.

To solve the aforementioned challenges, we propose a **G**lobal–**L**ocal **C**ontrastive **LE**arning **F**ramework (GL-CLEF) for zero-shot cross-lingual SLU. For the first challenge, as shown in Figure 1 (b), the key insight in GL-CLEF is to explicitly ensure that representations of similar sentences across languages are pulled closer together via contrastive learning (CL). Specifically, we leverage bilingual dictionaries to generate multi-lingual code-switched data pairs, which can be regarded as cross-lingual views with the same meaning. With the use of CL, our model is able to learn to distinguish the code-switched utterance of an input sentence from a set of negative examples, and thus encourages representations of similar sentences between source language and target language closer.

For the second challenge, SLU requires accomplishing tasks at two different levels: *token-level* slot filling and *sentence-level* intent detection. As such, simply leveraging ordinary *sentence-level* contrastive learning is ineffective for fine-grained knowledge transfer in *token-level* slot filling. Therefore, we first introduce a `Local` module in GL-CLEF to learn different granularity alignment representations (i.e., *sentence-level* `Local` intent CL and *token-level* `local` slot CL). To be specific, *sentence-level* `Local` intent CL and *token-level* `local` slot CL are introduced for aligning similar sentence and token representations across different languages for intent detection and slot filling, respectively. In addition, we further argue that slot and intent are highly correlated and have similar semantic meanings in a sentence. This phenomenon can serve as a signal for self-supervised alignment across intent and slots. Therefore, a `Global` module named *semantic-level* `global` intent–slot CL is further proposed to bring the representations of slot and intents within a sentence closer together.



(a) Training

(b) Zero-shot Testing

Figure 2: Zero-shot cross-lingual SLU.

We conduct experiments on MultiATIS++ (Xu et al., 2020), which includes nine different languages. Our experiments show that GL-CLEF achieves state-of-the-art results of 54.09% sentence accuracy, outperforming the previous best by 10.06% on average. Besides, extensive analysis experiments demonstrate that GL-CLEF has successfully reduced the representation gap between different languages.

To facilitate further research, codes are publicly available at `https://github.com/LightChen233/GL-CLeF`.

2 Background

We first describe traditional SLU before the specifics of zero-shot cross-lingual version of SLU.

Traditional SLU in Task-oriented Dialogue. SLU in Task-oriented Dialogue contains two subtasks: *Intent Detection* and *Slot Filling*.

· *Intent Detection:* Given input utterance $\mathbf{x}$, this is a classification problem to decide the corresponding intent label o^I.

· *Slot Filling:* Often modeled as a sequence labeling task that maps an input word sequence $\mathbf{x} = (x_1, \ldots, x_n)$ to slots sequence $\mathbf{o}^S = (o^S_1, \ldots, o^S_n)$, where n denotes the length of sentence $\mathbf{x}$.

Since the two tasks of intent detection and slot filling are highly correlated, it is common to adopt a joint model that can capture shared knowledge. We follow the formalism from Goo et al. (2018), formulated as $(o^I, o^S) = f(\mathbf{x})$, where f is the trained model.

Zero-shot Cross-lingual SLU. This means that a SLU model is trained in a source language, e.g., English (*cf.* Figure 2 (a)) and directly applied to other target languages (*cf.* Figure 2 (b)).

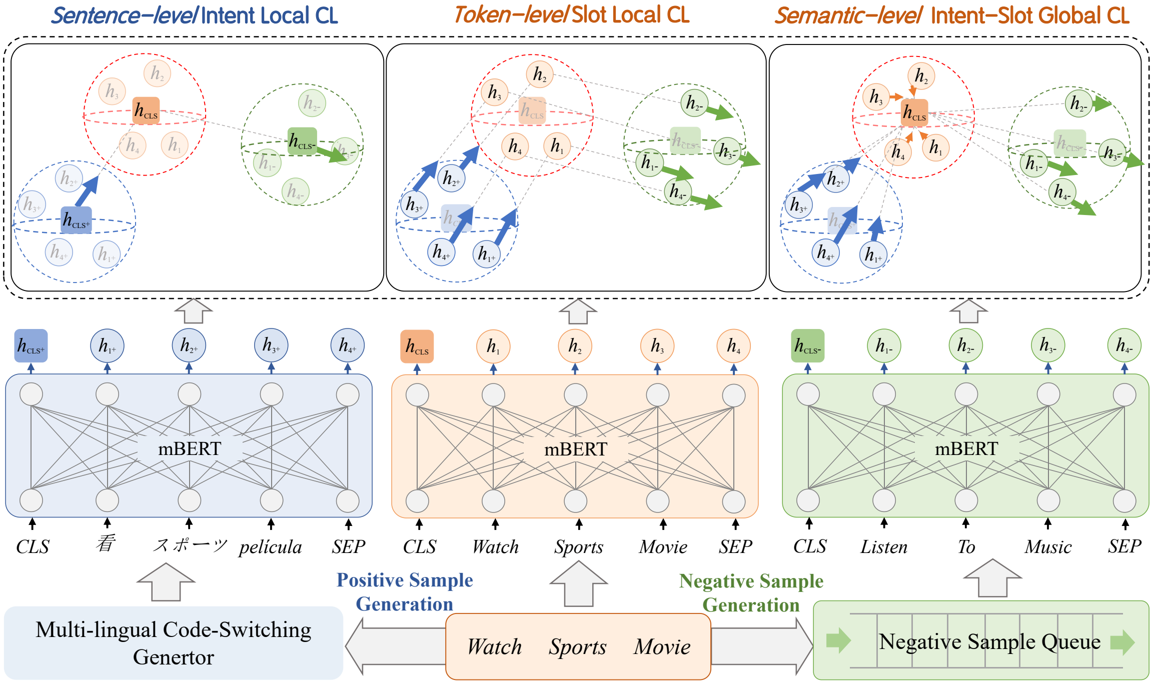

Sentence–level Intent Local CL
Token–level Slot Local CL
Semantic–level Intent–Slot Global CL
mBERT
mBERT
mBERT
CLS
看
スポーツ
película
SEP
CLS
Watch
Sports
Movie
SEP
CLS
Listen
To
Music
SEP
Positive Sample Generation
Negative Sample Generation
Multi-lingual Code-Switching Genertor
Watch Sports Movie
Negative Sample Queue

Figure 3: The main architecture of GL-CLEF. The boxes shown in figures are each sentence representation, while the circles are token representations. The dash lines and arrows in the top of the pictures on boxes and circles represent the direction of pushing in different levels made by contrastive learning. Different color denotes different representation spaces against anchor utterance, positive samples and negative samples. For simplicity, we only draw one case of *token-level* slot CL.

Formally, given each instance $\mathbf{x}_{tgt}$ in a target language, the model f which is trained on the source language is directly used for predicting its intent and slots:

$$(o^I_{tgt}, o^S_{tgt}) = f(\mathbf{x}_{tgt}), \tag{1}$$

where tgt represents the target language.

3 Model

We describe the general approach to general SLU task first, before describing our GL-CLEF model which explicitly uses contrastive learning to explicitly achieve cross-lingual alignment. The main architecture of GL-CLEF is illustrated in Figure 3.

3.1 A Generic SLU model

Encoder. Given each input utterance $x = (x_1, x_2, \ldots, x_n)$, the input sequence can be constructed by adding specific tokens $\mathbf{x} = (\texttt{[CLS]}, x_1, x_2, ..., x_n, \texttt{[SEP]})$, where `[CLS]` denotes the special symbol for representing the whole sequence, and `[SEP]` can be used for separating non-consecutive token sequences (Devlin et al., 2019). Then, we follow Qin et al. (2020) to first generate multi-lingual code-switched data. Then, we employ mBERT model to take code-switched data for encoding their representations $\mathbf{H} = (\boldsymbol{h}_{\texttt{CLS}}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_{\texttt{SEP}})$.

Slot Filling. Since mBERT produces subword-resolution embeddings, we follow Wang et al. (2019) and adopt the first sub-token's representation as the whole word representation and use the hidden state to predict each slot: $\boldsymbol{o}^S_t = \text{softmax}(\boldsymbol{W}^s \boldsymbol{h}_t + \boldsymbol{b}^s)$, where $\boldsymbol{h}_t$ denotes the first sub-token representation of word x_t; $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$ refer to the trainable parameters.

Intent Detection. We input the sentence representation $\boldsymbol{h}_{\texttt{CLS}}$ to a classification layer to find the label o^I: $o^I = \text{softmax}(\boldsymbol{W}^{\boldsymbol{I}} \boldsymbol{h}_{\texttt{CLS}} + \boldsymbol{b}^{\boldsymbol{I}})$, where $\boldsymbol{W}^{\boldsymbol{I}}$ and $\boldsymbol{b}^{\boldsymbol{I}}$ are tuneable parameters.

3.2 Global–local Contrastive Learning Framework

We introduce our global–local contrastive learning framework (GL-CLEF) in detail, which consists of three modules: 1) a *sentence-level* `local` intent contrastive learning (CL) module to align sentence representation across languages for intent detection, 2) a *token-level* `local` slot CL module to align token representations across languages for slot filling, and 3) *semantic-level* `global` intent–slot CL to align representations between a slot and an intent.

3.2.1 Positive and Negative Samples Construction

For contrastive learning, the key operation is to choose appropriate positive and negative pairs against to the original (anchor) utterance.

Positive Samples. Positive samples should preserve the same semantics compared against the anchor utterance. Therefore, given each anchor utterance $\mathbf{x} = (\texttt{[CLS]}, x_1, x_2, ..., x_n, \texttt{[SEP]})$, we follow Qin et al. (2020) to use bilingual dictionaries (Lample et al., 2018) to generate multi-lingual code-switched data, which is considered as the positive samples $\mathbf{x}_+$. Specifically, for each word x_t in $\mathbf{x}$, x_t is randomly chosen to be replaced with a translation provisioned from a bilingual dictionary to generate a positive sample. For example, given an anchor utterance *"watch sports movie"* in English, we can generate a positive multi-lingual code-switched sample *"看(watch/zh) スポツ (sports/ja) película (movie/es)"* (*cf.* Figure 3). Such a pair of anchor utterance and multi-lingual code-switched sample can be regarded as cross-lingual views of the same meaning across different languages. $\mathbf{x}_+$ is fed into mBERT to obtain the corresponding representations $\mathbf{H}_+ = (\boldsymbol{h}_{\texttt{CLS}+}, \boldsymbol{h}_{1+}, \ldots, \boldsymbol{h}_{n+}, \boldsymbol{h}_{\texttt{SEP}+})$.

Negative Samples. A natural approach for generating negative samples is randomly choosing other queries in a batch. However, this method requires the recoding of the negative samples, hurting efficiency. Inspired by He et al. (2020), in GL-CLEF, we maintain a negative sample queue, where the previously encoded original anchor utterance $\mathbf{x}$, positive samples $\mathbf{x}_+$ and previous negative samples $\mathbf{x}_-$ are also progressively reused as negative samples. This enables us to reuse the encoded samples from the immediate preceding batches, so as to eliminate the unnecessary negative encoding process. The negative sample queues for `[CLS]` and sentence representation are represented as: $\mathbf{H}_{\texttt{CLS}-} = \{\boldsymbol{h}^k_{\texttt{CLS}-}\}_{k=0}^{K-1}$, $\mathbf{H}_{S-} = \{\mathbf{H}^k_{S-}\}_{k=0}^{K-1}$, where K is the maximum capacity for negative queue.

3.2.2 `Local` Module

***Sentence-level* `Local` Intent CL.** Since intent detection is a *sentence-level* classification task, aligning sentence representation across languages is the goal of zero-shot cross-lingual intent detection task. Therefore, in GL-CLEF, we propose a *sentence-level* `local` intent CL loss to explicitly encourage the model to align similar sentence representations into the same local space across languages for intent detection. Formally, this is formulated as:

$$\mathcal{L}_{\mathrm{LI}} = -\log \frac{s(h_{\mathrm{CLS}}, h_{\mathrm{CLS^+}})}{s(h_{\mathrm{CLS}}, h_{\mathrm{CLS^+}}) + \sum_{\mathrm{k=0}}^{\mathrm{K\text{-}1}} s(h_{\mathrm{CLS}}, h^k_{\mathrm{CLS^-}})},$$

where $s(p, q)$ denotes the dot product between p and q; τ is a scalar temperature parameter.

***Token-level* `Local` Slot CL.** As slot filling is a token-level task, we propose a *token-level* `local` slot CL loss to help the model to consider token alignment for slot filling, achieving fine-grained cross-lingual transfer. We apply toke-level CL for all tokens in the query. Now, we calculate the ith token CL loss for simplicity:

$$\mathcal{L}^i_{\mathrm{LS}} = -\sum_{j=1}^{n} \log \frac{s(h_{\mathrm{i}}, h_{\mathrm{j^+}})}{s(h_{\mathrm{i}}, h_{\mathrm{j^+}}) + \sum_{k=0}^{\mathrm{K\text{-}1}} s(h_{\mathrm{i}}, h^k_{\mathrm{j^-}})} / n,$$

where the final $\mathcal{L}_{\mathrm{LS}}$ is the summation of all tokens CL loss.

3.2.3 `Global` Module

***Semantic-level* `Global` Intent-slot CL.** We noted that slots and intent are often highly related semantically when they belong to the same query. Therefore, we think that the intent in a sentence and its own slots can naturally constitute a form of positive pairings, and the corresponding slots in other sentences can form negative pairs. We thus further introduce a *semantic-level* `global` intent–slot CL loss to model the semantic interaction between slots and intent, which may further improve cross-lingual transfer between them. Formally:

$$\mathcal{L}_{\mathrm{GIS1}} = -\sum_{j=1}^{n} \log \frac{s(h_{\mathrm{CLS}}, h_{\mathrm{j}})}{s(h_{\mathrm{CLS}}, h_{\mathrm{j}}) + \sum_{k=0}^{\mathrm{K\text{-}1}} s(h_{\mathrm{CLS}}, h^k_{\mathrm{j^-}})} / n,$$

$$\mathcal{L}_{\mathrm{GIS2}} = -\sum_{j=1}^{n} \log \frac{s(h_{\mathrm{CLS}}, h_{\mathrm{j^+}})}{s(h_{\mathrm{CLS}}, h_{\mathrm{j^+}}) + \sum_{k=0}^{\mathrm{K\text{-}1}} s(h_{\mathrm{CLS}}, h^k_{\mathrm{j^-}})} / n,$$

$$\mathcal{L}_{\mathrm{GIS}} = \mathcal{L}_{\mathrm{GIS1}} + \mathcal{L}_{\mathrm{GIS2}},$$

where we consider CL loss from both anchor sentences ($\mathcal{L}_{\mathrm{GIS1}}$) and code-switched sentence ($\mathcal{L}_{\mathrm{GIS2}}$), and add them to do semantic-level contrastive learning ($\mathcal{L}_{\mathrm{GIS}}$) .

3.3 Training

3.3.1 Intent Detection Loss

$$\mathcal{L}_I \triangleq -\sum_{i=1}^{n_I} \hat{\mathbf{y}}_i^I \log\left(\mathbf{o}_i^I\right), \tag{2}$$

where $\hat{\mathbf{y}}_i^I$ are the gold intent label and n_I is the number of intent labels.

3.3.2 Slot Filling Loss

$$\mathcal{L}_S \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{n_S} \hat{\mathbf{y}}_j^{i,S} \log\left(\mathbf{y}_j^{i,S}\right), \tag{3}$$

where $\hat{\mathbf{y}}_j^{i,S}$ are the gold slot label for jth token; n_S is the number of slot labels.

3.3.3 Overall Loss

The overall objective in GL-CLeF is a tuned linear combination of the individual losses:

$$\mathcal{L} = \lambda_I\mathcal{L}_I + \lambda_S\mathcal{L}_S + \lambda_{\text{LI}}\mathcal{L}_{\text{LI}} + \lambda_{\text{LS}}\mathcal{L}_{\text{LS}} + \lambda_{\text{GIS}}\mathcal{L}_{\text{GIS}}, \tag{4}$$

where λ_* are tuning parameters for each loss component.

4 Experiments

We use the latest multilingual benchmark dataset of MultiATIS++ (Xu et al., 2020) which consists of 9 languages including English (en), Spanish (es), Portuguese (pt), German (de), French (fr), Chinese (zh), Japanese (ja), Hindi (hi), and Turkish (tr).

4.1 Experimental Setting

We use the base case multilingual BERT (mBERT), which has $N = 12$ attention heads and $M = 12$ transformer blocks. We select the best hyperparameters by searching a combination of batch size, learning rate with the following ranges: learning rate $\{2 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 6 \times 10^{-6}, 5 \times 10^{-5}, 5 \times 10^{-4}\}$; batch size $\{4, 8, 16, 32\}$; max size of negative queue $\{4, 8, 16, 32\}$; For all experiments, we select the best-performing model over the dev set and evaluate on test datasets. All experiments are conducted at TITAN XP and V100.

4.2 Baselines

To verify the effect of GL-CLeF, we compare our model with the following state-of-the-art baselines:
1) `mBERT`. mBERT[1] follows the same model architecture and training procedure as BERT (Devlin et al., 2019), but trains on the Wikipedia pages of 104 languages with a shared subword vocabulary. This allows mBERT to share embeddings across languages, which achieves promising performance on various cross-lingual NLP tasks;
2) `Ensemble-Net`. Razumovskaia et al. (2021) propose an `Ensemble-Net` where predictions are determined by 8 independent models through majority voting, each separately trained on a single source language, which achieves promising performance on zero-shot cross-lingual SLU;
3) `AR-S2S-PTR`. Rongali et al. (2020) proposed a unified sequence-to-sequence models with pointer generator network for cross-lingual SLU;
4) `IT-S2S-PTR`. Zhu et al. (2020) proposed a non-autoregressive parser based on the insertion transformer. It speeds up decoding and gain improvements in cross-lingual SLU transfer;
5) `CoSDA`. Qin et al. (2020) propose a data augmentation framework to generate multi-lingual code-switching data to fine-tune mBERT, which encourages the model to align representations from source and multiple target languages.

4.3 Main Results

Following Goo et al. (2018), we evaluate the performance of slot filling using F1 score, intent prediction using accuracy, and the sentence-level semantic frame parsing using overall accuracy which represents all metrics are right in an utterance.

From the results in Table 1, we observe that: (1) `CoSDA` achieves better performance than no alignment work `mBERT` and even outperforms the `Ensemble-Net`. This is because that such implicit alignment does align representations to some extent, compared against `mBERT`. (2) Our framework achieves the state-of-the art performance and beats `CoSDA` with 10.06% average improvements on overall accuracy. This demonstrates that `GL-CLeF` explicitly pull similar representations across languages closer, which outperforms the implicit alignment manner.

4.4 Analysis

To understand GL-CLeF in more depth, we perform comprehensive studies to answer the following research questions (RQs):
(1) Do the `local` intent and slot CLs benefit *sentence-* and *token-level* representation alignment? (2) Can *semantic-level* `global` intent-slot CL boost the overall sentence accuracy? (3) Are `local` intent CL and `local` slot CL complementary? (4) Does GL-CLeF pull similar representa-

[1] https://github.com/google-research/bert/blob/master/multilingual.md

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	AVG
mBERT* (Xu et al., 2020)	-	95.27	96.35	95.92	80.96	79.42	94.96	69.59	86.27	-
mBERT† (Devlin et al., 2019)	98.54	95.40	96.30	94.31	82.41	76.18	94.95	75.10	82.53	88.42
Ensemble-Net* (Razumovskaia et al., 2021)	90.26	92.50	96.64	95.18	77.88	77.04	95.30	75.04	84.99	87.20
CoSDA† (Qin et al., 2020)	95.74	94.06	92.29	77.04	82.75	73.25	93.05	80.42	78.95	87.32
GL-CLEF	**98.77**	**97.53**	**97.05**	**97.72**	**86.00**	**82.84**	**96.08**	**83.92**	**87.68**	**91.95**

Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	AVG
Ensemble-Net* (Razumovskaia et al., 2021)	85.05	82.75	77.56	76.19	14.14	9.44	74.00	45.63	37.29	55.78
mBERT* (Xu et al., 2020)	-	82.61	74.98	75.71	31.21	35.75	74.05	23.75	62.27	-
mBERT† (Devlin et al., 2019)	95.11	80.11	78.22	82.25	26.71	25.40	72.37	41.49	53.22	61.66
CoSDA† (Qin et al., 2020)	92.29	81.37	76.94	79.36	64.06	66.62	75.05	48.77	77.32	73.47
GL-CLEF	**95.39**	**86.30**	**85.22**	**84.31**	**70.34**	**73.12**	**81.83**	**65.85**	**77.61**	**80.00**

Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	AVG
AR-S2S-PTR* (Zhu et al., 2020)	86.83	34.00	40.72	17.22	7.45	10.04	33.38	–	23.74	-
IT-S2S-PTR* (Zhu et al., 2020)	87.23	39.46	50.06	46.78	11.42	12.60	39.30	–	28.72	-
mBERT† (Devlin et al., 2019)	87.12	52.69	52.02	37.29	4.92	7.11	43.49	4.33	18.58	36.29
CoSDA† (Qin et al., 2020)	77.04	57.06	46.62	50.06	26.20	28.89	48.77	15.24	46.36	44.03
GL-CLEF	**88.02**	**66.03**	**59.53**	**57.02**	**34.83**	**41.42**	**60.43**	**28.95**	**50.62**	**54.09**

Table 1: Results on MultiATIS++. We report both individual and average (AVG) test results on slot filling, intent detection accuracy, and overall accuracy. Results with "*" are taken from the corresponding published paper, while results with † are obtained by re-implemented. '–' denotes missing results from the published work.

tions across languages closer? (5) Does GL-CLEF improve over other pre-trained models? (6) Does GL-CLEF generalize to non pre-trained models? (7) Is GL-CLEF robust to the one-to-many translation problem?

Answer 1: `Local` intent CL and slot CL align similar sentence and token representations across languages. We investigate the effect of the `local` intent CL and `local` slot CL mechanism, by removing the `local` intent CL and slot CL, respectively (Figure 4, "– LI" and "– LS" (Col 1,2)). For the effectiveness of `local` intent CL, we find the performance of intent detection averaged on 9 languages drops by 3.52% against the full system (*ibid.* final, RHS column). This is because *sentence-level* intent CL loss can pull sentence representations closer across languages.

Similarly, considering the effectiveness of `local` slot CL, we find the performance of slot filling averaged on 9 languages drops by 2.44% against the full system. We attribute performance drops to the fact that `local` slot CL successfully make a fine-grained cross-lingual knowledge transfer for aligning token representation across languages, which is essential for *token-level* cross-lingual slot filling tasks.

Answer 2: *Semantic-level* `global` intent-slot successfully establishes a semantic connection across languages. We further investigate the effect of the *semantic-level* intent-slot CL mechanism when we remove the `global` intent-slot CL loss (Figure 4, "– GIS" (Col 3)). We find the sentence overall performance drops a lot (from 54.09% to 46.94%). Sentence overall metrics require model to capture the semantic information (intent and slots) for queries. Therefore, we attribute it to the proposed *semantic-level* `global` intent-slot CL. As it successfully establishes semantic connection across languages, it boosts overall accuracy.

Answer 3: Contribution from `local` intent CL and slot CL module are complementary. We explore whether `local` intent CL and slot CL module are complementary. By removing all the `Local` CL modules (including *sentence-level* `local` intent CL and *token-level* `local` slot CL), results are shown in Figure 4 (–Local Col 4). We find that the experiments are lowest compared with only removing any single `local` CL module, which demonstrates the designed two `local` CL module works orthogonally.

Answer 4: GL-CLEF pulls similar representations across languages closer. We choose test set and use representations of `[CLS]` of each sentence for visualization. Figure 5 (a, LHS) shows the t-SNE visualization of the `mBERT` output, where we observe that there very little overlap between different languages, which shows that the distance of the representations of different languages are distant. In contrast, the `GL-CLEF` representations (b, RHS) fine-tuned model in different languages are closer and largely overlap with each other. The stark contrast between the figures demonstrates

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
BiLSTM (Hochreiter and Schmidhuber, 1997)	72.56	70.96	70.35	60.05	64.50	64.33	71.75	56.22	60.13	65.65
BiLSTM+GL-CLEF	**84.77**	**74.44**	**71.09**	**69.53**	**65.29**	**66.14**	**77.02**	**63.36**	**67.08**	**70.97**
XLM-R (Conneau et al., 2020)	98.32	97.19	98.03	94.94	88.91	88.50	96.41	72.45	91.15	93.02
XLM-R+GL-CLEF	**98.66**	**98.43**	**98.04**	**97.85**	**93.84**	**88.83**	**97.76**	**81.68**	**91.38**	**94.05**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
BiLSTM (Hochreiter and Schmidhuber, 1997)	75.43	15.81	34.97	33.38	5.83	4.98	43.89	9.51	27.51	27.92
BiLSTM+GL-CLEF	**87.45**	**38.40**	**46.06**	**46.16**	**20.28**	**29.53**	**59.67**	**37.25**	**42.48**	**45.25**
XLM-R (Conneau et al., 2020)	94.58	72.35	76.72	71.81	60.51	9.31	70.08	45.21	13.44	57.38
XLM-R+GL-CLEF	**95.88**	**84.91**	**82.47**	**80.99**	**61.11**	**55.57**	**77.27**	**54.55**	**80.50**	**74.81**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	**AVG**
BiLSTM (Hochreiter and Schmidhuber, 1997)	37.06	0.78	3.08	0.63	0.22	0.00	10.20	0.00	0.03	5.80
BiLSTM+GL-CLEF	**61.37**	**4.60**	**9.10**	**4.30**	**0.34**	**2.03**	**16.82**	**2.80**	**2.46**	**11.53**
XLM-R (Conneau et al., 2020)	87.45	43.05	42.93	43.74	19.42	5.76	40.80	9.65	6.60	33.31
XLM-R+GL-CLEF	**88.24**	**64.91**	**53.51**	**58.28**	**19.49**	**13.77**	**52.35**	**14.55**	**52.07**	**46.35**

Table 2: Experimental results on BiLSTM and XLM-R.

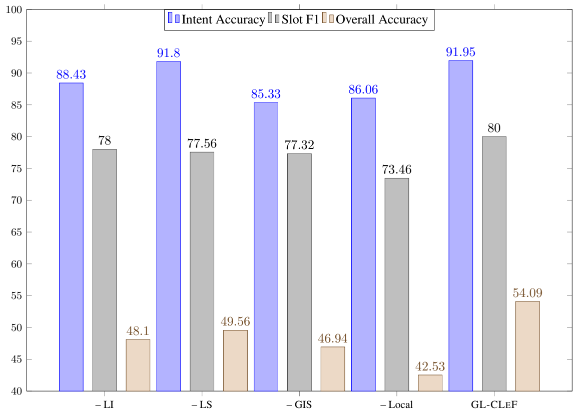
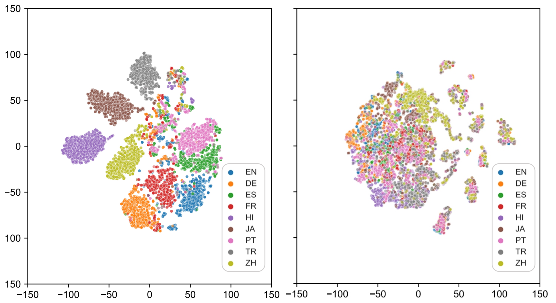


Figure 4: Ablation experiments. y-axis denotes the performance score. "LI", "LS" and "GIS" denote `Local` Intent CL, `Local` Slot CL, and `Global` Intent-Slot CL, respectively; "-Local" represents removing both "LI" and "LS" module.

that `GL-CLEF` successfully aligns representations of different languages.

Answer 5: Contributions from contrastive learning and pre-trained model use are complementary. To verify the contribution from GL-CLEF is still effective when used in conjunction with other strong pre-trained models, we perform experiments with `XLM-R` (Conneau et al., 2020). `XLM-R` demonstrates significant gains for a wide range of cross-lingual tasks. From the results in Table 2, we find GL-CLEF enhances `XLM-R`'s performance, demonstrating that contributions from the two are complementary. This also indicates that CL-CLEF is model-agnostic, hinting that GL-CLEF may be applied to other pre-trained models.

Answer 6: GL-CLEF still obtains gains over BiLSTM. A natural question that arises is



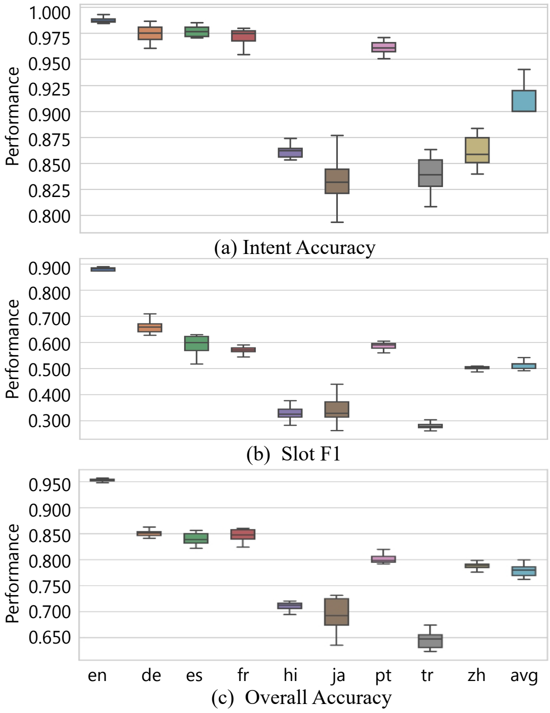
(a) Baseline

(b) GL-CLEF

Figure 5: t-SNE visualization of sentence vectors from (a) mBERT and (b) GL-CLEF. Different colors represents different languages.

whether `GL-CLEF` is effective for non pre-trained models, in addition to transformers. To answer the question, we replace `mBERT` with `BiLSTM`, keeping other components unchanged. The results are shown in Table 2. We can see that GL-CLEF outperforms `BiLSTM` in all metrics, further demonstrating that `GL-CLEF` is not only effective over `mBERT` but also ports to general encoders for both pre-trained models and non pre-trained models.

Answer 7: GL-CLEF is robust. It is worth noting that words in the source language can have multiple translations in the target language. We follow Qin et al. (2020) to randomly choose any of the multiple translations as the replacement target language word. Their work verified that random selection effective method (Qin et al., 2020). A natural question that arises is whether GL-CLEF is robust over different translation selections. To answer the question, we choose 15 different seeds to perform experiment and obtain the standard deviation, which we take as an indicator of the stability and robustness of models' performance. Results



(a) Intent Accuracy

(b) Slot F1

(c) Overall Accuracy

Figure 6: Performance distribution boxplots for each model over 15 random seeds.

in Figure 6 shows a lower standard deviation on each metric, indicating our model is robust to different translation. Finding and using the absolutely correct contextual word-to-word translation is an interesting direction to be explored in the future.

5 Related Work

Traditional Spoken Language Understanding. Since slot filling and intent detection are two correlated tasks, traditional SLU approaches mainly explore a joint model for capturing shared knowledge across the two tasks. Specifically, Zhang and Wang (2016); Liu and Lane (2016a,b); Hakkani-Tür et al. (2016) consider an implicit joint mechanism using a multi-task framework by sharing an encoder for both tasks. Goo et al. (2018); Li et al. (2018); Qin et al. (2019) consider explicitly leveraging intent detection information to guide slot filling. Wang et al. (2018); E et al. (2019); Zhang et al. (2020); Qin et al. (2021a) use a bi-directional connection between slot filling and intent detection.

Zero-shot Cross-lingual Spoken Language Understanding. Traditional SLU has largely been limited to high-resource languages. To solve this problem, zero-shot cross-lingual SLU has gained increasing attention. Recently, cross-lingual contextualized embeddings have achieved promising results (e.g., mBERT (Devlin et al., 2019)). Many works target improving mBERT at the pre-training stage (Conneau and Lample, 2019; Huang et al., 2019; Yang et al., 2020; Feng et al., 2020; Conneau et al., 2020; Xue et al., 2021; Chi et al., 2021a,b). Compared with their work, our focus is on enhancing mBERT at the fine-tuning stage.

In recent years, related work also considers aligning representations between source and target languages during fine-tuning, eschewing the need for an extra pre-training process. Specifically, Liu et al. (2020) propose code-mixing to construct training sentences that consist of both source and target phrases for implicitly fine-tuning mBERT. Qin et al. (2020) further propose a multi-lingual code-switching data augmentation to better align a source language and all target languages. In contrast to their work, our framework consider aligning similar representation across languages explicitly via a contrastive learning framework. In addition, in GL-CLEF, we propose a multi-resolution loss to encourage fine-grained knowledge transfer for token-level slot filling.

Contrastive Learning. Contrastive learning is now commonplace in NLP tasks. Wu et al. (2020) adopt multiple sentence-level augmentation strategies to learn a noise-invariant sentence representation. Fang and Xie (2020) apply back translation to create augmentations of original sentences for training transformer models. Wang et al. (2021) propose contrastive learning with semantically negative examples (CLINE) to improve the robustness under semantically adversarial attack. Inspired by the success of CL, we utilize contrastive learning to explicitly align similar representations across source language and target language.

6 Conclusion

We introduced a global–local contrastive learning (CL) framework (GL-CLEF) to explicitly align representations across languages for zero-shot cross-lingual SLU. Besides, the proposed `Local` CL module and `Global` CL module achieves to learn different granularity alignment (i.e., *sentence-level* local intent alignment, *token-level* local slot alignment, *semantic-level* global intent-slot alignment). Experiments on MultiATIS++ show that GL-CLEF obtains best performance and extensive analysis indicate GL-CLEF successfully pulls closer the representations of similar sentence across languages.

7 Ethical Considerations

Spoken language understanding (SLU) is a core component in task-oriented dialogue system, which becomes sufficiently effective to be deployed in practice. Recently, SLU has achieved remarkable success, due to the evolution of pre-trained models. However, most SLU works and applications are English-centric, which makes it hard to generalize to other languages without annotated data. Our work focuses on improving zero-shot cross-lingual SLU model that do not need any labeled data for target languages, which potentially is able to build multilingual SLU models and further promotes the globalization of task-oriented dialog systems.

Acknowledgements

We also thank all anonymous reviewers for their constructive comments. This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 62176078.

References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. of NAACL*.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021b. Xlm-e: cross-lingual language model pre-training via electra. *arXiv preprint arXiv:2106.16138*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proc. of NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proc. of ACL*.

Hongchao Fang and Pengtao Xie. 2020. CERT: contrastive self-supervised learning for language understanding. *ArXiv preprint*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *ArXiv preprint*.

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 *Proc. of NAACL*.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proc. of Interspeech*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. of CVPR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proc. of EMNLP*.

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

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proc. of EMNLP*.

Bing Liu and Ian Lane. 2016a. Attention-based recurrent neural network models for joint intent detection and slot filling.

Bing Liu and Ian Lane. 2016b. Joint online spoken language understanding and language modeling with recurrent neural networks. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

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 *Proc. of AAAI*.

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 *Proc. of EMNLP*.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021a. A co-interactive transformer for joint slot filling and intent detection. In *Proc. of ICASSP*.

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 *Proc. of IJCAI*.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021b. A survey on spoken language understanding: Recent advances and new frontiers. In *Proc. of IJCAI*.

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulić. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *ArXiv preprint*.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! A sequence to sequence architecture for task-oriented semantic parsing. In *Proc. of WWW*.

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

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive learning with semantic negative examples for natural language understanding. In *Proc. of ACL*.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proc. of NAACL*.

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*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: contrastive learning for sentence representation. *ArXiv preprint*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proc. of EMNLP*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL*.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *Proc. of AAAI*.

Linhao Zhang, Dehong Ma, Xiaodong Zhang, Xiaohui Yan, and Houfeng Wang. 2020. Graph LSTM with context-gated mechanism for spoken language understanding. In *Proc. of AAAI*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proc. of IJCAI*.

Qile Zhu, Haidar Khan, Saleh Soltan, Stephen Rawls, and Wael Hamza. 2020. Don't parse, insert: Multilingual semantic parsing with insertion based decoding. In *Proc. of CoNLL*.