Improving Zero-shot Cross-lingual Dialogue State Tracking via Contrastive Learning

Yu Xiang[1], **Ting Zhang**[2], **Hui Di**[3], **Hui Huang**[4], **Chunyou Li**[1], **Kazushige Ouchi**[3], **Yufeng Chen**[1], **Jinan Xu**[1*]

[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China

[2] Global Tone Communication Technology Co., Ltd.

[3] Toshiba (China) Co., Ltd. [4] Harbin Institute of Technology

{21120422,21120368,chenyf,jaxu}@bjtu.edu.cn;
zhangting01@gtcom.com.cn;dihui@toshiba.com.cn;
22b903058@stu.hit.edu.cn;kazushige.ouchi@toshiba.co.jp

Abstract

Recent works in dialogue state tracking (DST) focus on a handful of languages, as collecting large-scale manually annotated data in different languages is expensive. Existing models address this issue by code-switched data augmentation or intermediate fine-tuning of multilingual pre-trained models. However, these models can only perform implicit alignment across languages. In this paper, we propose a novel model named **C**ontrastive **L**earning for **C**ross-**L**ingual DST (CLCL-DST) to enhance zero-shot cross-lingual adaptation. Specifically, we use a self-built bilingual dictionary for lexical substitution to construct multilingual views of the same utterance. Then our approach leverages fine-grained contrastive learning to encourage representations of specific slot tokens in different views to be more similar than negative example pairs. By this means, CLCL-DST aligns similar words across languages into a more refined language-invariant space. In addition, CLCL-DST uses a significance-based keyword extraction approach to select task-related words to build the bilingual dictionary for better cross-lingual positive examples. Experiment results on Multilingual WoZ 2.0 and parallel MultiWoZ 2.1 datasets show that our proposed CLCL-DST outperforms existing state-of-the-art methods by a large margin, demonstrating the effectiveness of CLCL-DST.

1 Introduction

Dialogue state tracking is an essential part of task-oriented dialogue systems (Zhong et al., 2018), which aims to extract user goals or intentions throughout a dialogue process and encode them into a compact set of dialogue states, i.e., a set of slot-value pairs. In recent years, DST models have achieved impressive success with adequate training data. However, most models are restricted to monolingual scenarios since collecting and annotating task-oriented dialogue data in different languages is time-consuming and costly (Chen et al., 2018). It is necessary to investigate how to migrate a high-performance dialogue state tracker to different languages when no annotated target language dialogue data are available.

Previous approaches are generally divided into the following three categories: (1) Data augmentation methods with neural machine translation system (Schuster et al., 2019). Although translating dialogue corpora using machine translation is straightforward, it has inherent limitation of heavily depending on performance of machine translation. (2) Pre-trained cross-lingual representation (Lin and Chen, 2021). The approach applies a cross-lingual pre-trained model, such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) as one of the components of the DST architecture and then is trained with task data directly. However, the approach does not introduce cross-lingual information during the training process. (3) Code-switched data augmentation (Liu et al., 2020a; Liu et al., 2020b; Qin et al., 2021). The method replaces words randomly from the source language to the target language with a bilingual dictionary as a way to achieve data

* Corresponding author.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

augmentation. Nevertheless, a synonym substitution with some meaningless words may introduce noise that impairs the semantic coherence of the sentence. Besides, the model only use the code-switched corpus as the training data, ignoring the interaction between the original and code-switched sentences. Consequently, these models can not sufficiently learn the semantic representation of the corpus.

To address the above-mentioned issues, we propose a novel model named **C**ontrastive **L**earning for **C**ross-**L**ingual DST (CLCL-DST), which utilizes contrastive learning (CL) for cross-lingual adaptation. CLCL-DST first captures comprehensive cross-lingual information from different perspectives and explores the consistency of multiple views through contrastive learning (Lai et al., 2021). Simultaneously, as dialogue state tracking is to predict the state of slots in each turn of the dialogue, we consider it as a token-level task and then employ the same fine-grained CL. Specifically, we obtain the encoded feature representation of each slot in the original sentence and the corresponding code-switched sentence from the multilingual pre-trained model, respectively. We then employ fine-grained CL to align the representations of slot tokens in different views. By introducing CL, Our model is able to distinguish between the code-switched utterance and a set of negative samples, thus encouraging representations of similar words in different languages to align into a language-invariant feature space (Subsection 3.1).

Furthermore, CLCL-DST introduces a significance-based keyword extraction approach to obtain task-related keywords with high significance scores in different domains. For example, in the price range domain, some words like “cheap”, “moderate” and “expensive” are more likely to have higher significance scores than background words, such as “a”, “is” and “do”. Specifically, Our approach obtains the semantic representation of sentences and corresponding subwords by encoder. Then the approach gets the significance scores of the words by calculating the cosine similarity and get the keywords of the dataset based on the scores. We then replace these keywords with the corresponding words in the target language to generate multilingual code-switched data pairs. These code-switched keywords can be considered as cross-lingual views sharing the same meaning, allowing the shared encoder to learn some direct bundles of meaning in different languages. Thus, our keyword extraction approach facilitates the transfer of cross-lingual information and strengthens the ties across different languages (Subsection 3.2).

We evaluate our model on two benchmark datasets. For the Multilingual WoZ 2.0 dataset (Mrkšić et al., 2017) which is single-domain, our model outperforms the existing state-of-the-art model by 4.1% and 4.8% slot accuracy for German (De) and Italian (It) under the zero-shot setting, respectively. For the parallel MultiWoZ 2.1 dataset (Gunasekara, 2021) which is multi-domain, our method outperforms the current state-of-the-art by 22% and 38.7% in joint goal accuracy and slot f1 for Chinese (Zh), respectively. Moreover, further experiments show that introducing fine-grained CL performs better than coarse-grained CL. We also investigate the impact of different keyword extraction approaches on the model to demonstrate the superiority of our extraction approach.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work on DST that leverages fine-grained contrastive learning to explicitly align representations across languages.
- We propose to utilize a significance-based keyword selection approach to select task-related keywords for code-switching. By constructing cross-lingual views through these keywords makes the model more effective in transferring cross-lingual signals.
- Our CLCL-DST model achieves state-of-the-art results on single-domain cross-lingual DST tasks, and it boasts the unique advantage of performing effective zero-shot transfer under the multi-domain cross-lingual setting, demonstrating the effectiveness of CLCL-DST.

2 Related Work

2.1 Dialogue State Tracking

Methods of dialogue state tracking can be divided into two categories, ontology-based and open-vocabulary DST. The first method selects the possible values for each slot directly from a pre-defined ontology and the task can be seen as a value classification task for each slot (Lee et al., 2019; Goel et

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

al., 2019; Lin et al., 2021; Wang et al., 2022). However, in practical applications, it is difficult to define all possible values of slots in advance, and the computational complexity increases significantly with the size of the ontology.

The open-vocabulary approach attempts to solve the above problems by extracting or generating slot values directly from the dialogue history (Ren et al., 2019). (Wu et al., 2019) generates slot values directly for each slot at every dialogue turn. The model uses GRU to encode the dialogue history and decode the value with a copy mechanism. Some recent works (Kim et al., 2020; Zeng and Nie, 2020b) adopt a more efficient approach by decomposing DST into two tasks: state operation prediction and value generation. SOM-DST (Kim et al., 2020) firstly predicts state operation on each slot and then generates the value of the slot that needs updating. (Zeng and Nie, 2020a) proposes a framework based on the architecture of SOM-DST, with a single BERT as both the encoder and the decoder.

2.2 Zero-shot Cross-Lingual Dialogue State Tracking

There is a growing demand for dialogue systems supporting different languages, which requires large-scale training data with high quality. However, these data are only available within a few languages. It remains a challenge to migrate dialogue state tracker from the source language to the target language.

Cross-lingual dialogue state tracking can be divided into two categories: single-domain and multi-domain. In single-domain, XL-NBT (Chen et al., 2018) first implements cross-lingual learning under the zero-shot setting by pre-training a dialogue state tracker for the source language using a teacher network. MLT (Liu et al., 2020a) adopts a code-mixed data augmentation framework, leveraging attention mechanism to obtain the code-mixed training data for learning the interlingual semantics across different languages. CLCSA (Qin et al., 2021) further explores the dynamic replacement of words from source language to target language during training. Based on CLCSA architecture, XLIFT-DST (Moghe et al., 2021) improves the performance by intermediate fine-tuning of pre-trained multilingual models using parallel and conversational movie subtitles datasets.

In multi-domain, the primary benchmark is the Parallel MultiWoZ 2.1 dataset (Gunasekara, 2021) originating from the Ninth Dialogue Systems and Technologies Challenge (DSTC-9) (Gunasekara, 2021). This challenge is designed to build a dialogue state tracker to evaluate a low-resource target language dataset using the learned knowledge of the source language. All the submissions in this challenge use the translated version of the dataset, transforming the problem into a monolingual dialogue state tracking task. XLIFT-DST employs SUMBT (Lee et al., 2019) as the base architecture and achieves competitive results on the parallel MultiWoZ 2.1 dataset through intermediate fine-tuning. Unlike these works, we leverage code-switched data with CL to further align multiple language representations under the zero-shot setting.

2.3 Contrastive Learning

Contrastive learning aims at pulling close semantically similar examples (positive samples) and pushing apart dissimilar examples (negative samples) in the representation space. SimCSE (Gao et al., 2021) proposes a simple dropout approach to construct positive samples and achieves state-of-the-art results in semantic textual similarity tasks. Cline (Wang et al., 2021) constructs semantically negative instances without supervision to improve the robustness of the model against semantically adversarial attacks. GL-CLEF (Qin et al., 2022) leverages bilingual dictionaries to generate code-switched data as positive samples, and incorporates different grained contrastive learning to achieve cross-lingual transfer. Our model incorporates fine-grained CL to align similar representations between the source and target languages.

3 Methodology

In this section, we set up the notations that run throughout the paper first, before describing our CLCL-DST model which explicitly uses contrastive learning to achieve cross-lingual alignment in dialogue state tracking. Then, we introduce a significance-based code-switching approach on how to select task-related keywords in the utterance and code-switch the input sentence dynamically in detail. The main

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

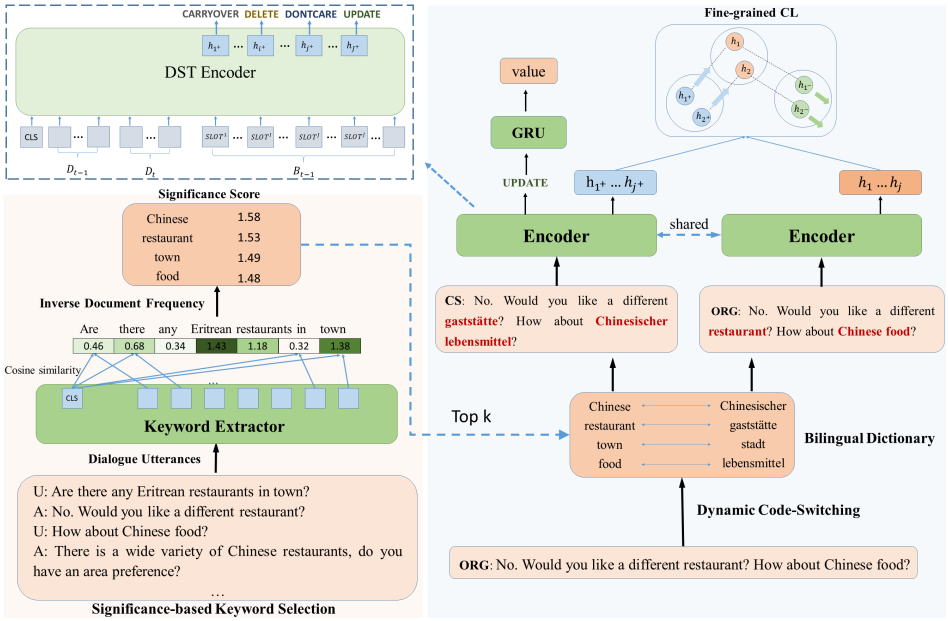

Figure 1: The overview of the proposed CLCL-DST. The input of our model consists of previous turn dialogue utterances D_{t-1}, current turn dialogue utterances D_t and previous dialogue state B_{t-1}. For simplicity, we only put one turn of dialogue on the picture. The model constructs a bilingual dictionary by obtaining keywords from the significance-based code-switching approach, and then generates code-switched data. The data are fed to the encoder to obtain a feature representation of each slot subsequently. **ORG** denotes the original sentence and **CS** denotes the corresponding code-switched sentence. In the part of **Fine-grained CL**, different color denotes different representation spaces for origin utterance, positive and negative samples. The decoder generates the value for the slot whose state operation is predicted to **UPDATE**.

architecture of our model is illustrated in Figure 1.

Notation. Suppose the dialogue has T turns. We define the dialogue utterance at turn t as $D_t = R_t \oplus ; \oplus U_t \oplus [\text{SEP}]$, where R_t and $U_t (1 \leq t \leq T)$ are the system response and the user utterance respectively. $\oplus$ denotes token concatenation, and the semicolon ; is a separation symbol, while [SEP] marks the end boundary of the dialogue. Besides, we represent the dialogue states as $B = \{B_1, ..., B_T\}$, where $B_t = [\text{SLOT}]^1 \oplus b_t^1 \oplus ... \oplus [\text{SLOT}]^{\text{I}} \oplus b_t^I$ denotes I states combination at the t-th turn. I is the total number of slots. The i-th slot-value pair b_t^i is defined as:

$$b_t^i = S^i \oplus - \oplus V_t^i, \tag{1}$$

where S^i is a slot and V_t^i is the corresponding slot value. $[\text{SLOT}]^{\text{i}}$ and $-$ are separation symbols. The representations at $[\text{SLOT}]^{\text{i}}$ position are used for state operation prediction and contrastive learning. We use the same special token [SLOT] for all $[\text{SLOT}]^{\text{i}}$. The input tokens in CLCL-DST are spliced by previous turn dialogue utterance D_{t-1}, current turn dialogue utterance D_t and previous turn dialogue states B_{t-1} (Kim et al., 2020):

$$X_t = [\text{CLS}] \oplus D_{t-1} \oplus D_t \oplus B_{t-1}, \tag{2}$$

where [CLS] is a special token to mark the start of the context. Next, we will elaborate each part in detail.

3.1 Fine-grained Contrastive Learning Framework

We introduce our fine-grained contrastive learning framework (CLCL-DST) with an encoder-decoder architecture consisting of two modules: state operation prediction and value generation. The encoder, i.e., state operation predictor, uses a multilingual pre-trained model to predict the type of the operations

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

to be performed on each slot. The decoder, i.e., slot value generator, generates values for those selected slots.

Encoder The encoder of CLCL-DST is based on mBERT architecture. We feed the code-switched sentence $X_{t,cs}$ into the encoder and obtain the output representation $H_{t,cs} \in \mathbb{R}^{|X_t| \times d}$, where $h_{t,cs}^{[\mathrm{CLS}]}, h_{t,cs}^{[\mathrm{SLOT}]^i} \in \mathbb{R}^d$ are the outputs corresponding to [CLS] and $[\mathrm{SLOT}]^i$. $h_{t,cs}^{[\mathrm{SLOT}]^i}$ is passed into a four-way classification layer to calculate the probability $P_{enc,t}^i \in \mathbb{R}^{|\mathcal{O}|}$ of operations in the i-th slot at the t-th turn:

$$P_{enc,t}^i = \mathrm{softmax}\left(W_{enc} h_{t,cs}^{[\mathrm{SLOT}]^i} + b\right), \tag{3}$$

where W_{enc} and b are learnable parameters. $\mathcal{O} = \{\mathrm{CARRYOVER}, \mathrm{DELETE}, \mathrm{DONTCARE}, \mathrm{UPDATE}\}$ denotes four state operations of each slot (Kim et al., 2020). Specifically, CARRYOVER indicates that the slot value remains unchanged; DELETE changes the value to NULL; and DONTCARE means that the slot is not important at this turn and does not need to be tracked (Wu et al., 2019). Only when the UPDATE is predicted does the decoder generate a value for the corresponding slot.

Our main learning objective is to train the encoder to match predicted state operation with the ground truth operation. So the loss for state operation is formulated as:

$$\mathcal{L}_{enc,t} = -\frac{1}{I} \sum_{i=1}^{I} \left(Y_{enc,t}^i\right)^\top \log\left(P_{enc,t}^i\right), \tag{4}$$

where $Y_{enc,t}^i \in \mathbb{R}^{|\mathcal{O}|}$ is the ground truth operation for the j-th slot.

Decoder We employ GRU as decoder to generate the value of dialogue state for each domain-slot pair whose operation is UPDATE. GRU is initialized with $g_t^{i,0} = W_t$ and $e_t^{i,0} = h_t^{[\mathrm{SLOT}]^i}$. The probability distribution of the vocabulary is calculated as:

$$P_{dec,t}^{i,k} = \mathrm{softmax}\left(\mathrm{GRU}\left(g_t^{i,k-1}, e_t^{i,k}\right) \times E\right) \in \mathbb{R}^{|V|}, \tag{5}$$

where k is decoding step, $E \in \mathbb{R}^{|V| \times d}$ is the word embedding space shared with the encoder, and $|V|$ is the size of multilingual vocabulary. The overall loss for generating slot value is the average of the negative log-likelihood loss:

$$\mathcal{L}_{dec,t} = -\frac{1}{|\mathbb{U}_t|} \sum_{i \in \mathbb{U}_t} \left[\frac{1}{K_t^i} \sum_{k=1}^{K_t^i} \left(Y^{i,k}\right)^\top \log\left(P_{dec,t}^{i,k}\right)\right], \tag{6}$$

where $|\mathbb{U}_t|$ is the number of slots which require value generation, K_t^i indicates the number of ground truth value to be generated for the i-th slot. $Y^{i,k} \in \mathbb{R}^{|V|}$ represents the one-hot vector of the ground truth token generated for the i-th slot at the k-th decoding step.

Fine-grained Contrastive Learning In order to better capture the common features between the source language and the target language, our model utilizes fine-grained CL to pull closer the representation of similar sentences across different languages. The key to CL is to find high-quality positive and negative pairs corresponding to the original utterance. The positive sample should be semantically consistent with the original utterance and provides cross-lingual view as well. In our scenario, we choose code-switched input $X_{t,cs}$ as the positive sample of X_t, while other inputs in the same batch are treated as negative samples.

As state operation of each slot is a token-level task, we utilize a fine-grained CL loss to facilitate token alignment. To achieve fine-grained cross-lingual transfer, our method selects the output representation $h_t^{[\mathrm{SLOT}]^i}$ of the special token $[\mathrm{SLOT}]^i$ for contrastive learning, as these I tokens are able to convey the semantics of the slots in the query. The i-th slot token loss is defined as:

$$\mathcal{L}_{cl,t}^i = -\frac{1}{I} \sum_{j=1}^{I} \log \frac{cos\left(h_t^i, h_t^{j^+}\right)}{cos\left(h_t^i, h_t^{j^+}\right) + \sum_{k=0,k\neq j}^{\mathrm{I}-1} cos\left(h_t^i, h_t^{k^-}\right)}, \tag{7}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

where h_t^i is the abbreviation of $h_t^{[\text{SLOT}]^i}$, $h_t^{j^+}$ and $h_t^{k^-}$ are positive and negative samples of $h_t^{[\text{SLOT}]^i}$ respectively. The total loss $\mathcal{L}_{cl,t}$ is calculated by adding up all tokens CL loss.

The overall objective in CLCL-DST at dialogue turn t is the sum of individual losses above:

$$\mathcal{L}_t = \mathcal{L}_{enc,t} + \mathcal{L}_{cl,t} + \mathcal{L}_{dec,t}. \tag{8}$$

3.2 Significance-based Code-switching

The importance of different words in a dialogue utterance varies. For example, in the price range domain, "cheap" and "expensive" are more likely to be keywords, while in the area domain, keyword set might include orientation terms such as "center", "north" and "east". Assuming that a dataset contains v words constituting a vocabulary $\mathcal{V}$, we construct a subset of keywords $\mathcal{K} \subseteq \mathcal{V}$ for code-switching. Subsequently, the encoder of CLCL-DST serves to extract keywords in the training data.

Given the input token $X_t = (w_t^1, w_t^2, ..., w_t^n)$ at the t-th turn, n denotes the number of words. We feed X_t into encoder, and obtain the representation $h_t^{[\text{CLS}]} \in \mathbb{R}^d$ of the special token [CLS]. Then the sentence embedding vector W_t is calculated as:

$$W_t = \tanh(W_{pool} h_t^{[\text{CLS}]} + b), \tag{9}$$

where W_{pool} and b are learnable parameters. Then the cosine similarity between each token $w_t \in X_t$ and the sentence embedding vector W_t is computed as:

$$Sim(w_t) = \cos(w_t, W_t). \tag{10}$$

$Sim(w_t)$ reflects the degree of associations between w_t and sentence embedding W_t. A higher value of the significance score $Sim(w_t)$ indicates a higher probability of w_t to be a keyword. For words that are tokenized into subwords, we average the significance scores of each subword to obtain the word score.

Equation 10 calculates the significance score of words in a sentence. To get the keyword set $\mathcal{K}$ in training set, we add all significance scores for token w in training set and multiply them by the inverse document frequency (IDF) (Yuan et al., 2020) of w:

$$S(w) = \log \frac{N}{|\{x \in X : w \in x\}|} \cdot \sum_{x \in X : w \in x} Sim(w), \tag{11}$$

where N denotes the number of the input in the training dataset, $|\{x \in X : w \in x\}|$ indicates the number of the input containing w. The IDF term can reduce the weight of words which appear frequently in the dataset, assigning meaningless words (e.g., "for" and "an") with a lower score.

We select top-k words according to the significance scores to get a keyword set K, and use the bilingual dictionary MUSE (Lample et al.,) to construct the code-switched dictionary $Dic = ((s_1, t_1), ...(s_k, t_k))$, where s and t refer to the source and target language words respectively. k is the number of keywords. In addition, we translate the whole words in ontology and add them to Dic due to their important role in the sentence.

Inspired by (Qin et al., 2021), we randomly replace some words in source language sentence with corresponding target words with a fixed probability if they appear in Dic. Since words from the source language may have multiple translations in Dic, we randomly select one of them for substitution. Notably, the input token X in our model includes dialogue utterance D and dialogue states B, we just replace source words in D as B shares the same slots across languages. Finally, we can get the code-switched input tokens $X_{t,cs}$ from X_t as:

$$X_{t,cs} = [\text{CLS}] \oplus D_{t-1,cs} \oplus D_{t,cs} \oplus B_{t-1}, \tag{12}$$

4 Experiments

4.1 Datasets

We evaluate our model on two datasets as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

- **Multilingual WoZ 2.0 dataset** (Mrkšić et al., 2017): A restaurant domain dialogue dataset expanded from WoZ 2.0 (Wen et al., 2017), which contains three languages (English, German, Italian) and 1200 dialogues for each language. The corpus consists of three goal-tracking slot types: food, price range and area. The task is to learn a dialogue state tracker only in English and evaluate it on the German and Italian datasets, respectively.
- **Parallel MultiWoZ dataset** (Gunasekara, 2021): A seven domains dialogue dataset expanded from MultiWoZ 2.1 (Eric et al., 2020). Parallel MultiWoZ contains two languages (English, Chinese) and 10K dialogues. The Chinese corpus is obtained through Google Translate and manually corrected by experts.

4.2 Compared Methods

We compare our approach with the following methods:

- **XL-NBT** (Chen et al., 2018) utilizes bilingual corpus and bilingual dictionaries to transfer the teacher's knowledge of the source language to a student tracker in the target languages.
- **MLT** (Liu et al., 2020a) constructs code-switched data through the attention layer for training.
- **CLCSA** (Qin et al., 2021) dynamically constructs multilingual code-switched data by randomly replacing words, so as to better fine-tune mBERT and achieve outstanding results in multiple languages.
- **SUMBT** (Lee et al., 2019) uses a non-parametric distance measure to score each candidate slot-value pair. We replace BERT with mBERT on the cross-lingual setup.
- **SOM-DST** (Kim et al., 2020) employs BERT as the encoder and uses a copy-based RNN to decode upon BERT outputs.
- **DST-as-PROMPTING** (Lee et al., 2021) introduces an approach that uses schema-driven prompting to provide history encoding and then utilizes T5 to generate slot values directly. Here, we use the multilingual version of T5 - mT5 (Xue et al., 2021).
- **XLIFT-DST** (Moghe et al., 2021) leverages task-related parallel data to enhance transfer learning by intermediate fine-tuning of pre-trained multilingual models. For parallel MultiWoZ, XLIFT-DST uses the architecture of SUMBT, while uses the state tracker in CLCSA for Multilingual WoZ 2.0.

4.3 Implementation Details

Our method leverages the pre-trained mBERT-base[0] implemented by HuggingFace as the encoder, with 12 Transformer blocks and 12 self-attention heads. One layer GRU is used as the decoder. The encoder shares the same hidden size s with the decoder, which is 768. Adam optimizer (Kingma and Ba, 2014) is applied to optimize all parameters with a warmup strategy for the 10% of the total training steps. The peak learning rate is set to 4e-5 for encoder and 1e-4 for decoder, respectively. Besides, we use greedy decoding for generating slot values.

For Multilingual WoZ dataset, the batch size is set to 64 and the maximum sequence length to 200. For parallel MultiWoZ dataset, the batch size and the maximum sequence length are 16 and 350 respectively. We replace the word for each dialogue with a fixed probability of 0.6. The training is performed for 100 epochs as default, and we choose the best checkpoint on the validation set to test our model.

4.4 Evaluation Metrics

The metrics in dialogue state tracking are turn-level which include Slot Accuracy, Joint Goal Accuracy and Slot F1. Slot Accuracy is the proportion of the correct slots predicted in all utterances. Joint Goal Accuracy is the proportion of dialogue turns where all slot values predicted at a turn exactly match the ground truth values, while Slot F1 is the Macro-average of F1 score computed over the slot values at each turn.

[0]https://huggingface.co/bert-base-multilingual-uncased

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

Model	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
XL-NBT (Chen et al., 2018)	55.0	30.8	72.0	41.2
MLT (Liu et al., 2020a)	69.5	32.2	69.5	31.4
Transformer based				
mBERT	57.6	15.0	54.6	12.6
CLCSA (Qin et al., 2021)	83.0	63.2	82.2	61.3
XLIFT-DST (Moghe et al., 2021)	85.2	**65.8**	84.3	66.9
CLCL-DST (ours)	**89.3**	63.2	**89.1**	**67.0**

Table 1: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset under zero-shot setting when trained with English task data. Please see text for more details. **Bold** indicates the best score in that column. CLCL-DST denotes our approach.

Model	joint acc.	slot f1.
SUMBT (Lee et al., 2019) †	1.9	14.8
SOM-DST (Kim et al., 2020) ‡	1.7	10.6
DST-as-PROMPTING (Lee et al., 2021) ‡	2.5	17.6
XLIFT-DST †	5.1	40.7
CLCL-DST (ours)	**27.1**	**79.4**
In-language training †	15.8	70.2
Translate-Train †	11.1	54.2
Translate-Test †	26.5	77.0

Table 2: Joint goal accuracy and slot F1 on parallel MultiWoZ dataset under zero-shot learning setting when trained with English task data and tested on Zh language. '†' denotes results from (Moghe et al., 2021). '‡' denotes our re-implemented results for the models based on corresponding multilingual pretrained models.

4.5 Main Results

Results for the Multilingual WoZ dataset are illustrated in Table 1. We can see that CLCL-DST outperforms the state-of-the-art model (XLIFT-DST) by 4.1% and 4.8% in slot accuracy for De and It respectively. This demonstrates that our model is able to explicitly bring similar representations of different languages closer together through contrastive learning than augmenting transfer learning process with intermediate fine-tuning of pre-trained multilingual models.

To further study the effectiveness of our model under the zero-shot setting, We also test CLCL-DST on parallel MultiWoZ in Table 2. As there are only a few baselines available for this dataset, we re-implement some monolingual models such as SUMBT, SOM-DST, DST-as-PROMPTING into multilingual scenarios. We find that our model has 22% and 38.7% improvement over XLIFT-DST in joint goal accuracy and slot f1 for target language Zh under the zero-shot setting. It is worth noting that the joint goal accuracy of all these baseline models is relatively low. The possible reason is that these models do not learn considerable cross-lingual representations in the multi-domain cases, making it difficult to migrate for complex slots. Specifically, In the SOM-DST model, its decoder utilizes the soft-gated copy mechanism (See et al., 2017) in addition to GRU, which introduces additional noise from the source language and is not applicable to multilingual settings. In DST-as-PROMPTING, the model only leverages mT5 to generate slot values directly without learning deeply cross-lingual interaction information. Besides, we also refer to the results of translation-based methods from (Moghe et al., 2021) in Table 2. Our model still outperforms all of them. These results further indicate that our proposed CLCL-DST leveraging code-switched data with contrastive learning boosts the performance of dialogue state tracker.

5 Ablation Studies

We conduct ablation experiments to explore the effect of fine-grained contrastive learning and the significance-based keyword extraction approach on the overall performance for the Multilingual WoZ

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

2.0 dataset.

5.1 The Effect of Fine-grained Contrastive Learning

In addition to fine-grained CL, we also introduce coarse-grained CL for aligning similar sentences across different languages. To be specific, we align the sentence embedding W_t from equation 9 with its corresponding code-switched positive representations W_t^+. The objective for coarse-grained CL is written as follows:

$$\mathcal{L}_{sl,t} = -\log \frac{\cos\left(W_t, W_t^+\right)}{\cos\left(W_t, W_t^+\right) + \sum_{k=0,k\neq j}^{\mathrm{I}-1} \cos\left(W_t, W_t^{k-}\right)}, \tag{13}$$

where $W_t^{k^-}$ is the negative sample for W_t at the t-th turn.

Method	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
w/o CL	82.5	52.0	86.8	60.0
Coarse-grained CL	87.9	57.7	79.8	41.0
Fine-grained CL	**89.3**	**63.2**	**89.1**	**67.0**

Table 3: Slot accuracy and joint goal accuracy for different grained contrastive learning under zero-shot setting. "CL" denotes the abbreviation of contrastive learning.

As results shown in Table 3, we can conclude that different granularities of contrastive learning are effective for our model, especially fine-grained CL since it can bring more improvement to CLCL-DST. Using fine-grained CL improves 1.4% and 5.5% in slot accuracy and joint goal accuracy for De, and 9.3% and 26% for It, respectively, compared to coarse-grained CL. Since the goal of dialogue state tracking is to predict the state of slots in each turn of the dialogue, it can be considered as a token-level task, so fine-grained CL is better suited for this task compared to coarse-grained CL. Also, our approach selects specific tokens representing slots instead of all tokens in the dialogue for contrastive learning, which can reduce the noise caused by other semantically irrelevant tokens.

5.2 The Effect of significance-based code-switching

In this section we further explore the impact of keyword extraction algorithm on CLCL-DST. Table 4 shows the performance of different keyword extraction strategies. We try other four approaches to obtain the mapping dictionaries and compare them with the significance-based code-switching approach: (1) choosing words based on their frequency in our training set and converting them to target languages by MUSE; (2) using the whole ontology, which contains 90 words approximately; (3) combining the dictionaries obtained from (1) and (2) to form a new dictionary; (4) extracting keywords using only TF-IDF algorithm.

Method	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
MUSE	86.4	59.4	84.0	54.5
Onto	86.2	56.0	81.8	46.8
MUSE+Onto	88.0	57.8	88.4	66.3
TF-IDF+Onto	86.5	55.3	87.9	66.0
Significance-based	87.9	60.4	**89.1**	63.5
Significance-based+Onto	**89.3**	**63.2**	**89.1**	**67.0**

Table 4: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset for different keywords extraction approaches under zero-shot setting. The Method column represents the strategy for extracting keywords. "Onto" is the abbreviation of ontology. "+" denotes the merging of dictionaries obtained by the two methods.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.

Number of keywords	German		Italian	
	slot acc.	joint acc.	slot acc.	joint acc.
200	86.5	60.4	85.1	61.5
500	88.2	62.3	86.8	64.4
1000	**89.3**	63.2	**89.1**	**67.0**
2000	88.6	**63.3**	86.9	66.3
5000	88.9	62.9	87.4	66.5

Table 5: Slot accuracy and joint goal accuracy on Multilingual WoZ 2.0 dataset for different number of keywords under zero-shot setting.

Compared with only considering the frequency of words in the corpus, our significance-based code-switching approach can also make use of the numerous linguistic information carried in the multilingual pretrained model, so that the selected words are more representative of the utterance. This approach enables the selected words to better express the main idea of the text. At the same time, words in ontology such as place names, food names, etc. are originally special words in the dataset, which occupy an important position in the text. Adding these words to our dictionary can further improve the performance of the model.

Table 5 shows the influence of different number of keywords on our model. We can see that the model has the best or second-best performance when k is 1000. As k continues to increase, the additional keywords are less indicative, so they even have a negative impact on model performance.

6 Conclusion

In this paper, we propose a novel zero-shot adaptation method CLCL-DST for cross-lingual dialogue state tracking. Our approach leverages fine-grained contrastive learning to explicitly align representations across languages. Besides, we introduce the significance-based code-switching approach to replace task-relevant words with target language for generating code-switched sentences on downstream tasks. Our method obtains new state-of-the-art results on Multilingual WoZ dataset and parallel MultiWoZ dataset, which demonstrates its effectiveness. In the future, we would investigate better training objectives for cross-lingual DST task, especially on multi-domain area, to further boost the dialogue system on multi-lingual scenarios. We would also explore better positive and negative samples when applying contrastive learning on DST task.

Acknowledgement

The research work descried in this paper has been supported by the National Key RD Program of China (2020AAA0108005), the National Nature Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130) and Toshiba (China) Co., Ltd. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 414–424.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL).

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *Proc. Interspeech 2019*, pages 1458–1462.

Chulaka Gunasekara. 2021. Overview of the ninth dialog system technology challenge: Dstc9. In *DSTC9 Workshop at AAAI 2021*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582.

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

Siyu Lai, Hui Huang, Dong Jing, Yufeng Chen, Jinan Xu, and Jian Liu. 2021. Saliency-based multi-view mixed language training for zero-shot cross-lingual classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 599–610.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.

Yen-Ting Lin and Yun-Nung Chen. 2021. An empirical study of cross-lingual transferability in generative dialogue state tracker. *arXiv preprint arXiv:2101.11360*.

Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021. Knowledge-aware graph-enhanced gpt-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020a. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Zihan Liu, Genta Indra Winata, Peng Xu, Zhaojiang Lin, and Pascale Fung. 2020b. Cross-lingual spoken language understanding with regularized representation alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7241–7251.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1137–1150.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

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.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. 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 1876–1885.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. Cline: Contrastive learning with semantic negative examples for natural language understanding. 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 2332–2342.

Yifan Wang, Jing Zhao, Junwei Bao, Chaoqun Duan, Youzheng Wu, and Xiaodong He. 2022. Luna: Learning slot-turn alignment for dialogue state tracking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3319–3328.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Michelle Yuan, Mozhi Zhang, Benjamin Van Durme, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996.

Yan Zeng and Jian-Yun Nie. 2020a. Jointly optimizing state operation prediction and value generation for dialogue state tracking. *arXiv preprint arXiv:2010.14061*.

Yan Zeng and Jian-Yun Nie. 2020b. Multi-domain dialogue state tracking based on state graph. *arXiv preprint arXiv:2010.11137*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 624-635, Harbin, China, August 3 - 5, 2023.