A Coarse-to-Fine Prototype Learning Approach for Multi-Label Few-Shot Intent Detection

Xiaotong Zhang[1], **Xinyi Li**[1], **Feng Zhang**[2], **Zhiyi Wei**[1], **Junfeng Liu**[1], **Han Liu**[*1]
[1]School of Software Technology, Dalian University of Technology
[2]School of Computer Science, Peking University
{zhangxt,hanliu}@dlut.edu.cn, {lixinyi_dlut,zhiyi.wei,junfeng.liu}@hotmail.com,
zfeng.maria@gmail.com

Abstract

Few-shot intent detection is a challenging task, particularly in scenarios involving multiple labels and diverse domains. This paper presents a novel prototype learning approach that combines the label synset augmentation and the coarse-to-fine prototype distillation for multi-label few-shot intent detection. To tackle the data scarcity issue and the lack of information for unseen domains, we propose to enhance the representations of utterances with label synset augmentation and refine the prototypes by distilling the coarse domain knowledge from a universal teacher model. To solve the multilingual intent detection in real-world dialogue systems, we fine-tune a cross-lingual teacher model to make our method fast adapt to different languages and re-annotate two non-English task-oriented dialogue datasets CrossWOZ and JMultiWOZ in multi-label form. Experimental results on one English and two non-English datasets demonstrate that our approach significantly outperforms existing methods in terms of accuracy and generalization across different domains.

1 Introduction

Intent detection which aims to identify intents behind user utterances is a core component of task-oriented dialogue systems (Shen et al., 2021; Chen et al., 2017), as its performance directly affects downstream decisions and policies. In real-world conversation scenarios, a single utterance could contain multiple intents, and the ways of expressing an intent are diverse. The data scarcity issue makes intent detection rather challenging as user intents constantly emerge in rapidly changing domains (Vulić et al., 2022). Therefore, the imperative to accurately recognize multiple intents in the low-data regime motivates the multi-label few-shot intent detection (FS-MLID).

Existing works mainly focus on the popular metric-based meta-learning paradigm for the FS-MLID task, which aims to learn a metric space that can make label predictions by calculating distances between query samples and prototypes of different classes. By training on a set of sampled FS-MLID tasks, the model learns general knowledge to rapidly generalize to new tasks with novel intent classes. In particular, CTLR (Hou et al., 2021) proposes to estimate label-instance relevance scores and uses a meta-calibrated threshold to select multiple associated intent labels. DCKPN (Zhang et al., 2023) constructs a dual class knowledge propagation network that combines label information and feature structure to guide intent prediction.

However, existing methods neglect that it is difficult to estimate the class prototypes in low-resource settings, and they also lack domain information to predict novel classes (Wang et al., 2024), thereby diminishing the discriminability of the metric space and model generalization. In addition, previous works merely focus on monolingual setting and only conduct on English datasets (Khalil et al., 2019). In contrast to English, most other languages lack sufficient annotated data to train high-quality intent detection models, which will ultimately hinder the application of task-oriented dialogue systems to a much wider spectrum of languages.

To address the aforementioned issues, we revisit the FS-MLID task from a multilingual perspective and propose a novel **C**oarse-to-**F**ine **P**rototype **L**earning method (CFPL), which is shown in Figure 1. Considering the scarcity of samples in few-shot learning and the rich semantic information beneath class labels, we propose to enhance the representations of utterances with label synset augmentation. Specifically, we first generate a synset for each intent label using Open Multilingual Wordnet (Bond et al., 2016), then we propose a refinement method to further eliminate the noise in the expanded label set. To precisely estimate the class prototypes

[*]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2489–2502
November 12-16, 2024 ©2024 Association for Computational Linguistics

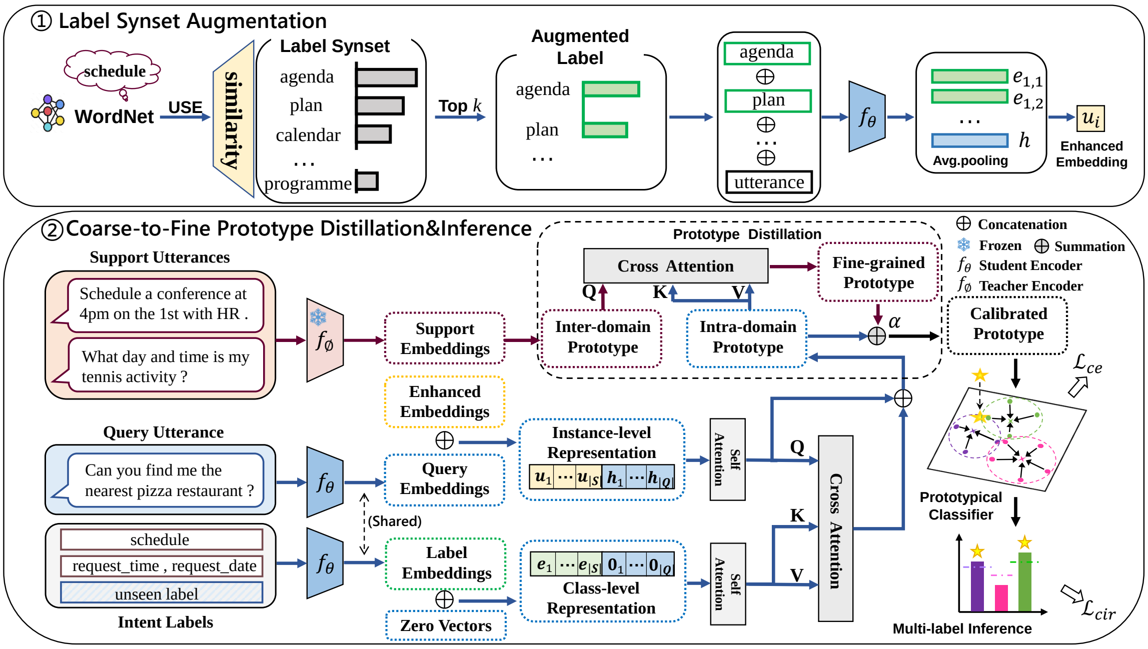


Figure 1: The framework of the CFPL method. First, a label synset is generated from WordNet, using semantic similarity to select the top augmented labels for each class. These augmented labels are then concatenated with support utterances to form enhanced embeddings. In the coarse-to-fine prototype distillation and inference stage, the enhanced embeddings of support data, the query embeddings, and the label embeddings are combined to learn instance-level and class-level representations for the support and query data in each episode. Then these representations are interacted to obtain the intra-domain prototypes via self-attention and cross-attention. Meanwhile, the inter-domain prototypes are learned by fine-tuning a cross-lingual teacher model, and are further distilled into fine-grained prototypes via cross-attention. Finally, the intra-domain and fine-grained prototypes are fused into calibrated prototypes, which are input into a prototypical classifier to predict multiple intents for each query data.

in the absence of domain knowledge, we further devise the coarse-to-fine prototype distillation. In particular, intra-domain student prototypes are first learned through feature interactions at the instance and class levels for a specific dataset. Then the inter-domain teacher prototypes that contain coarse domain knowledge from the teacher model are distilled into fine-grained prototypes for a specific dataset, which are further fused with intra-domain student prototypes to constitute more precise prototypes. Moreover, we fine-tune a cross-lingual teacher model to make our method fast adapt to different languages. To verify this, we re-annotate two non-English task-oriented dialogue datasets in multi-label form, i.e., CrossWOZ (Zhu et al., 2020) and JMultiWOZ (Ohashi et al., 2024), which contain multiple domains and thus can simulate the few-shot scenario in unseen domains. The contributions of this paper can be summarized as follows:

(1) We propose a coarse-to-fine prototype learning approach to recognize multiple intents for an utterance in low-resource settings. We first design a label augmentation strategy with semantic similarity refinement to generate enhanced data representations. During the prototype distillation, we first conduct feature interactions between samples at the instance and class levels to learn intra-domain student prototypes, then we distill related coarse domain knowledge from the universal teacher model into fine-grained prototypes for a specific dataset.

(2) To bridge the multilingual gap, we propose a simple but efficient fine-tuning method that enables the teacher model to fast adapt to different languages. Furthermore, we introduce and release two non-English FS-MLID datasets, which is an important attempt towards multilingual intent detectors for task-oriented dialogues.

(3) Extensive experiments demonstrate that our proposed methods outperform competitive baselines on three FS-MLID benchmarks, and is adept at handling low-resource situations.

2 Related Works

2.1 Multi-Label Intent Detection

Intent detection aims to mine the main purpose behind user utterances. Many studies (Goo et al., 2018; Qin et al., 2019; Liu et al., 2021) have achieved promising performance for intent detection. However, they neglect the more practical and challenging scenario, multi-label intent detection, which aims to assign multiple intents to samples. Rychalska et al. (2018) firstly propose to conduct multi-label intent detection. Considering the close relationship between intent detection and slot filling, Gangadharaiah and Narayanaswamy (2019), Qin et al. (2020) and Qin et al. (2021) design different strategies to leverage slot information to enhance multi-intent detection. Zhu et al. (2024) introduce the global static and local dynamic heterogeneous label graph to model interactions among samples. Wu et al. (2021) propose to construct a label embedding space by using label words. Vulic et al. (2022) conduct contrastive conversational fine-tuning on pre-trained sentence encoders.

2.2 Multi-Label Few-Shot Learning

Few-shot learning (FSL) aims to learn from limited labeled samples and recognize novel classes that have not been seen during the training process. Compared with single-label FSL, multi-label FSL is more common in many real scenarios, but only a few works have been done. Previous works focus on image domain (Alfassy et al., 2019) or audio domain (Cheng et al., 2019). In natural language processing domain, Liu et al. (2022) propose to address the multi-label aspect category detection task with a novel label-enhanced prototypical network. Only a few studies such as CTLR (Hou et al., 2021) and DCKPN (Zhang et al., 2023) have addressed the FS-MLID scenario, but they cannot well solve the data scarcity issue and the lack of information for unseen domains. Moreover, they only focus on monolingual setting.

3 The Proposed Method

3.1 Problem Definition

Few-shot learning aims to train a model that can recognize unknown categories with few labeled examples (Snell et al., 2017). In accordance with prior works, we follow the episodic paradigm on account of its effectiveness (Yang et al., 2021). Given a set of training classes $\mathcal{C}_{train}$ and testing classes $\mathcal{C}_{test}$, where $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$. The model is trained with numerous samples from $\mathcal{C}_{train}$, then directly adopted to unseen classes $\mathcal{C}_{test}$ with few labeled samples. In each episode, we have a support set $\mathcal{S} = \{(x_i, \boldsymbol{y_i})\}_{i=1}^{N \times K}$ where x_i represents a data sample, $\boldsymbol{y_i}$ is the corresponding class label, N is the number of classes and K is the number of support data in each class, and a query set $\mathcal{Q} = \{(x_j, \boldsymbol{y_j})\}_{j=1}^{Q}$ where $\mathcal{Q}$ is the number of query samples.

Multi-label few-shot intent detection allows that each utterance is associated with multiple intents. Given an utterance x, its label can be represented with a vector $\boldsymbol{y} = [y^1, y^2, ..., y^N]$, where $y^i \in \{0, 1\}$ and N is the number of possible intents.

3.2 Framework Overview

Our method consists of three components: Label Synset Augmentation, Coarse-to-Fine Prototype Distillation, and Optimization and Inference. We begin with Label Synset Augmentation to conduct label augmentation for each class and enhance the representations of support data with these augmented labels. Then we implement Coarse-to-Fine Prototype Distillation, which involves four parts. Inter-Domain Prototype Learning applies a pre-trained teacher model to capture coarse domain features. In Cross-Lingual Teacher section, a cross-lingual teacher model is fine-tuned to adapt quickly to different languages. Intra-Domain Prototype Learning conducts feature interactions for the support and query data within each episode. Prototype Distillation produces fine-grained prototypes from the coarse inter-domain prototypes and combines it with intra-domain student prototypes to get the final prototypes. During the Optimization and Inference, we adopt a prototypical classifier and multi-label inference to train the whole model and achieve multi-intent prediction for each query data.

3.3 Label Synset Augmentation

Many prior methods on data augmentation have validated the effectiveness of label enhancement in the few-shot learning setting (Luo et al., 2021; Zhang et al., 2022; Liu et al., 2022). For intent detection, the core issue is to extract user intents related to the utterance from all aspects and granularities. Hence, we choose Open Multilingual Wordnet (Bond et al., 2016), a large lexical database of synsets for over 150 languages, to generate multiple synonyms related to each original label. The process of label synset augmentation is shown in Figure 1. For each training label y_i, we generate a set of augmented labels $\mathcal{Y}_i = \{\tilde{y}_{i,1}, \tilde{y}_{i,2}, \cdots \tilde{y}_{i,n}\}$. Different from previous research that only uses Synonym Replacement (Wei and Zou, 2019) which randomly selects an augmented label to replace the original intent label, we further propose a label refinement method to eliminate the noisy labels from the set $\mathcal{Y}_i$.

An ideal augmented label $\tilde{y}$ should have high semantic similarity with the original label y as well as the corresponding user utterance x (Hu et al., 2022). Thus we concatenate the utterance and its original label into a sequence s as a basis for selecting enhanced labels. We introduce an auxiliary semantic similarity calculation function $sim(\, . \, , . \,)$ to guide the selection, where $sim(\, . \, , . \,)$ is a model that can output the semantic similarity between two text samples. In particular, for a sequence s, we use multilingual USE (Chidambaram et al., 2019) to compute the similarity between s and each augmented label, and select the top k augmented labels to enhance the representations for support data.

$$\mathcal{J} = \text{top-}k([sim(s, \tilde{y}_{i,j})]_{j=1}^{n}), \tag{1}$$

where top-$k(\mathcal{A})$ returns the indices of the largest k elements of the set $\mathcal{A}$, and k is a hyperparameter.

Denoting $\{\tilde{y}_{i,j}\}_{j \in \mathcal{J}}$ as a subset of k augmented labels from $\mathcal{Y}_i$, we can add them before the utterance and obtain the label enhanced embedding for each utterance x_i through the student model:

$$\boldsymbol{u}_i = \text{AvgPooling}([\boldsymbol{e}_1; \cdots; \boldsymbol{e}_j; \boldsymbol{h}_{x_i}]), \tag{2}$$

where ; represents concatenation along the token length dimension, and AvgPooling$(\cdot)$ is to obtain a vector by performing average pooling on a set of vectors along the token length dimension. $\boldsymbol{e}_j$ is the embedding of the label $\tilde{y}_{i,j}$, and $\boldsymbol{h}_{x_i}$ is the feature vector of the utterance x_i, which are all obtained from the student model $f_\theta(\cdot)$ such as Bert.

Finally, we learn the label enhanced representations of all the support data through the student model.

3.4 Coarse-to-Fine Prototype Distillation

In this section, we propose to learn the prototypes in each episode with knowledge distillation. Knowledge distillation algorithms aim to exploit the hidden knowledge from a large teacher model, denoted as T, to guide the training of a small student model, denoted as S (Hinton et al., 2015). Different from conventional distillation that teacher provides soft-targets for students, we propose to distill features to help the student to leverage the semantic knowledge of related domains from the teacher.

Inter-Domain Prototype Learning To pre-train a teacher model, we adopt the multilingual-BERT as the backbone, and train it on the Massive dataset (FitzGerald et al., 2023) in English which covers 60 intent classes from 18 domains.

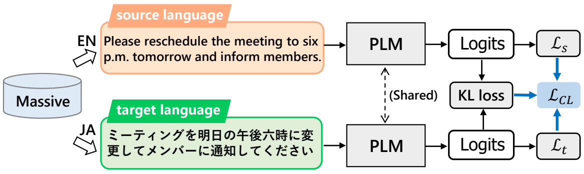

source language
EN
Please reschedule the meeting to six p.m. tomorrow and inform members.
Massive
target language
JA
ミーティングを明日の午後六時に変更してメンバーに通知してください
PLM
PLM
(Shared)
Logits
Logits
KL loss
$\mathcal{L}_s$
$\mathcal{L}_{CL}$
$\mathcal{L}_t$

Figure 2: Fine-tuning cross-lingual teacher.

Then for each utterance x_i in the support set, we learn its representation through the teacher encoder f_ϕ:

$$\boldsymbol{g}_i = f_\phi(x_i), \tag{3}$$

where $\boldsymbol{g}_i$ is the state vector corresponding to the $[CLS]$ token. Based on its label y_i, we can construct an inter-domain teacher prototype, which incorporates related semantic information from 18 domains.

$$\boldsymbol{p}_{\mathcal{T}}^{c} = \frac{1}{|\mathcal{S}_c|} \sum_{y_i = c} \boldsymbol{g}_i, \tag{4}$$

where c is the intent category of the given domain, and $|\mathcal{S}_c|$ is the number of support data belonging to category c in the support set.

Cross-Lingual Teacher The vast majority of previous methods focus on developing models on English datasets (Wu et al., 2022; Chen et al., 2017), which can hardly adapt to the datasets of other languages unless they train a model from scratch for a target language. In this paper, we propose to fine-tune the teacher model f_ϕ so that it can fast adapt to different languages, which is shown in Figure 2. In particular, given a pre-trained teacher model in the source language such as English, we add an extra model parameter $\boldsymbol{W}_{lm}$ to fine-tune the teacher model to convert it from the source language to the target language. The predicted soft labels or logits of x_i in the source language and target language are calculated by:

$$\begin{aligned} \hat{\boldsymbol{y}}_i^s &= \text{Softmax}(\boldsymbol{W}_{lm}\boldsymbol{g}_i^s), \\ \hat{\boldsymbol{y}}_i^t &= \text{Softmax}(\boldsymbol{W}_{lm}\boldsymbol{g}_i^t), \end{aligned} \tag{5}$$

where $\boldsymbol{g}_i^s$ and $\boldsymbol{g}_i^t$ are the embeddings of an utterance x_i in the source language and target language, respectively. We aim to align the labels of each support data in different languages, so that the teacher model in the source language could convert to the one in the target language.

$$\mathcal{L}_{KL} = \frac{1}{N}\sum_{i=1}^{N}(KL(\hat{\boldsymbol{y}}_i^s||\hat{\boldsymbol{y}}_i^t) + KL(\hat{\boldsymbol{y}}_i^t||\hat{\boldsymbol{y}}_i^s)), \quad (6)$$

where $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss to enforce the label probability distributions in the source and target languages similar. Moreover, we hope the predicted labels are close to the ground truth:

$$\mathcal{L}_s = \text{CrossEntropy}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^s), \quad (7)$$

$$\mathcal{L}_t = \text{CrossEntropy}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i^t). \quad (8)$$

Finally, we obtain the overall loss for fine-tuning the cross-lingual teacher model:

$$\mathcal{L}_{CL} = \mathcal{L}_{KL} + \mathcal{L}_s + \mathcal{L}_t. \quad (9)$$

Intra-Domain Prototype Learning The inter-domain teacher prototypes learned through Eq. (4) can capture the high-level semantic knowledge from 18 domains, which incorporate the related inter-domain information into the prototype of each class. But these teacher prototypes ignore to utilize the data information of each episode (which is from a single domain), thus we further propose to learn the intra-domain prototypes according to the support and query data within each episode.

Intuitively, if the data with the same label have similar representations, the prediction will become easier and more precise. Therefore in this section, we hope to further use attention mechanism to perform message passing among the data of each episode, so that the support embeddings and query embeddings could interact with each other.

Firstly, we merge the label enhanced representations of the support set and the original representations of the query set, and obtain the instance-level representations for the data of each episode:

$$\boldsymbol{\mathcal{I}} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{|\mathcal{S}|}, \boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{|\mathcal{Q}|}], \quad (10)$$

where $\boldsymbol{u}_i$ is the label enhanced support embedding, and $\boldsymbol{h}_i$ is the query embedding obtained by the original utterance. $|\mathcal{S}|$ and $|\mathcal{Q}|$ are the sizes of support and query sets in each episode.

Similarly, according to the labels of support data, we merge the label embeddings of the data in each episode, and obtain the class-level representations:

$$\boldsymbol{\mathcal{C}} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{|\mathcal{S}|}, \boldsymbol{0}_1, \boldsymbol{0}_2, \cdots, \boldsymbol{0}_{|\mathcal{Q}|}], \quad (11)$$

where $\boldsymbol{e}_i = f_\theta(y_i)$ is the label embedding for the label y_i. Since the labels for the query data are unknown, we represent the label embeddings of query data with a zero vector $\boldsymbol{0}$.

Then we perform self-attention on instance-level representations and class-level representations respectively to achieve information interaction within each episode:

$$\boldsymbol{A}^I = \text{Softmax}\left(\frac{(\boldsymbol{\mathcal{I}}\boldsymbol{W}_Q^1)(\boldsymbol{\mathcal{I}}\boldsymbol{W}_K^1)^T}{\sqrt{d_h}}\right)(\boldsymbol{\mathcal{I}}\boldsymbol{W}_V^1), \quad (12)$$

$$\boldsymbol{A}^C = \text{Softmax}\left(\frac{(\boldsymbol{\mathcal{C}}\boldsymbol{W}_Q^2)(\boldsymbol{\mathcal{C}}\boldsymbol{W}_K^2)^T}{\sqrt{d_h}}\right)(\boldsymbol{\mathcal{C}}\boldsymbol{W}_V^2), \quad (13)$$

where $\boldsymbol{A}^I$ and $\boldsymbol{A}^C$ are the interacted instance-level representations and class-level representations of the data in each episode, respectively.

We further use cross-attention operations to select the most related class-level representations according to the instance-level representations:

$$\hat{\boldsymbol{A}}^C = \text{Softmax}\left(\frac{(\boldsymbol{A}^I\boldsymbol{W}_Q^3)(\boldsymbol{A}^C\boldsymbol{W}_K^3)^T}{\sqrt{d_h}}\right)(\boldsymbol{A}^C\boldsymbol{W}_V^3). \quad (14)$$

We fuse the instance-level representations with the aligned class-level representations through concatenation operations and obtain the representations of all the data in each episode:

$$\boldsymbol{A} = \boldsymbol{A}^I||\hat{\boldsymbol{A}}^C. \quad (15)$$

Finally, we use the fused data representations $\boldsymbol{A}$ to construct the intra-domain student prototype for each class c:

$$\boldsymbol{p}_{\mathcal{S}}^c = \frac{1}{|\mathcal{S}_c|}\sum_{y_i=c}\boldsymbol{A}_i. \quad (16)$$

Prototype Distillation To this end, we propose to distill the most representative inter-domain teacher prototypes into fine-grained prototypes with cross-attention:

$$\boldsymbol{p}'_{\mathcal{T}} = \text{Softmax}\left(\frac{(\boldsymbol{p}_{\mathcal{T}}\boldsymbol{W}_Q^4)(\boldsymbol{p}_{\mathcal{S}}\boldsymbol{W}_K^4)^T}{\sqrt{d_h}}\right)(\boldsymbol{p}_{\mathcal{S}}\boldsymbol{W}_V^4), \quad (17)$$

where $\boldsymbol{p}_{\mathcal{S}}$ and $\boldsymbol{p}_{\mathcal{T}}$ are the intra-domain student prototype matrix and the inter-domain teacher prototype matrix, respectively. Then we combine it with intra-domain student prototypes to get the final prototypes for each episode:

$$\boldsymbol{p} = \alpha\boldsymbol{p}_{\mathcal{S}} + (1-\alpha)\boldsymbol{p}'_{\mathcal{T}}, \quad (18)$$

where α is a trade-off hyperparameter ranging between 0 and 1.

3.5 Optimization and Inference

Prototypical Classifier Given a query utterance $x_i \in \mathcal{Q}$, we can compute the conditional probability $p(y = c|x_i, \mathcal{S})$ to predict its labels based on negative squared Euclidean distance.

$$p(y = c|x_i, \mathcal{S}) = \frac{\exp(-||\boldsymbol{a}_i - \boldsymbol{p}_c||_2^2)}{\sum_{c' \in \mathcal{C}} \exp(-||\boldsymbol{a}_i - \boldsymbol{p}_{c'}||_2^2)}, \tag{19}$$

where $\boldsymbol{p}_c$ is the prototype of class c from $\boldsymbol{p}$, and $\boldsymbol{a}_i$ is the representation of x_i from $\boldsymbol{A}$.

Note that in the multi-label setting, as an utterance may have multiple labels, we need to consider $|N|$ labels for each query sample. We perform cross-entropy loss on all the query samples, which is calculated as:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{Q}|} \sum_{x_i \in \mathcal{Q}} \sum_{j=1}^{|N|} y_i^j \text{log} p(y = j|x_i, \mathcal{S}), \tag{20}$$

where y_i = $\{y_i^1, ..., y_i^{|N|}\}$ is the label of x_i and y_i^j $\in \{0, 1\}$.

Multi-Label Inference Inspired by (Sun et al., 2020), we introduce the class-specific circle loss to conduct multi-label prediction for each query sample x_i:

$$\begin{aligned}\mathcal{L}_{cir} =& \frac{1}{N} \sum_{c=1}^{N} (\log(e^{\sigma(\tau_c)} + \sum_{z_i \in \Lambda_c} e^{\sigma(z_i)}) \\ &+ \log(e^{\sigma(-\tau_c)} + \sum_{z_j \in \Gamma_c} e^{\sigma(-z_j)})),\end{aligned} \tag{21}$$

where σ is the temperature scale parameter, $\Lambda_c = \{p(y = c|x_i, \mathcal{S})|y_i^c = 0\}$ is the negative score set, $\Gamma_c = \{p(y = c|x_i, \mathcal{S})|y_i^c = 1\}$ is the positive score set, and τ_c is the threshold of class c. The goal of $\mathcal{L}_{cir}$ is that the positive scores of class c are greater than τ_c and the negative scores of class c are less than τ_c.

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cir}, \tag{22}$$

where $\lambda \in (0, 1)$ is a trade-off hyperparameter.

4 Experiments

4.1 Datasets and Experimental Setups

Datasets We follow (Zhang et al., 2023) to evaluate our method on public English FS-MLID dataset StanfordLU and introduce two new non-English FS-MLID datasets, i.e., CrossWOZ and JMultiWOZ. **(1) StanfordLU** is a dataset of Stanford dialogues (Eric et al., 2017), which contains 8038 utterances re-annotated by (Hou et al., 2021) from 3 domains: Sc (Schedule), Na (Navigate) and We (Weather). **(2) CrossWOZ** is an re-annotated version of the first large-scale Chinese multi-domain task-oriented dialogue dataset (Zhu et al., 2020) containing 8697 user utterances and includes 3 domains: At (Attraction), Ho (Hotel) and Re (Restaurant). **(3) JMultiWOZ** consists of 8076 utterances and includes three travel-related domains: Sh (Shopping), Ho (Hotel) and Re (Restaurant), which is re-annotated from the first Japanese multi-domain task-oriented dialogue dataset JMultiWOZ (Ohashi et al., 2024). We re-annotate the two publicly available non-English datasets into multi-label form and follow (Hou et al., 2020) to construct few-shot episodes. For each dataset, we take two domains as the training set and validation set respectively, and take another domain as the test set. We construct 200, 50, and 50 episodes for training, validation and testing, respectively. Table 1 shows detailed dataset statistics [1].

Dataset	StanfordLU			CrossWOZ			JMultiWOZ		
Domain	Sc	Na	We	At	Ho	Re	Sh	Ho	Re
Ns	14	10	8	8	10	9	8	10	10
Prop.	21%	25%	4%	48%	48%	53%	45%	60%	79%

Table 1: Dataset statistics. Ns denotes the number of classes in each domain and Prop. denotes the proportion of multi-label utterances.

Implementation Details The proposed approach CFPL is implemented with PyTorch and all the experiments are conducted on NVIDIA GeForce RTX 3090. In terms of feature extraction, we use bert-base-uncased, bert-base-Chinese and bert-base-Japanese as the student model respectively, and we use multilingual-BERT (mBERT) as the teacher model (Devlin et al., 2019). The size of the hidden state is 768 and the number of hidden layers is 12. We use AdamW (Loshchilov and Hutter, 2019) for optimization with the initial learning rate of 2e-5 on StanfordLU, 1e-4 on CrossWOZ, and 5e-5 on JMultiWOZ. We set the dropout rate as 0.2, hyperparameter α as 0.8, σ as 0.05 and the number of augmented labels k as 2 (detailed analysis is shown in 4.5). For the loss function, we set λ as 0.1. All the hyperparameters are determined by

[1] The source code and data are available at https://github.com/CFPL2024/CFPL

Models	StanfordLU 1-shot				StanfordLU 5-shot			
	Sc	Na	We	Avg.	Sc	Na	We	Avg.
TransferM	18.00±0.62	24.65±0.79	22.26±0.64	21.64±0.68	16.62±0.18	23.69±0.46	26.64±2.04	22.31±0.89
MMN	39.18±0.52	35.35±1.72	45.87±2.81	40.13±1.68	43.65±6.24	51.94±1.03	46.65±0.48	47.41±2.58
MPN	39.34±1.38	36.09±0.77	45.86±2.50	40.43±1.55	41.45±2.83	50.51±2.94	54.96±9.76	48.97±5.18
CTLR	42.55±0.40	56.95±0.77	53.14±1.89	50.88±1.02	52.17±1.29	60.36±1.55	59.63±2.23	57.39±1.69
DCKPN	53.81±0.72	58.48±0.31	74.02±0.74	62.10±0.59	57.81±0.62	63.71±0.35	**93.83±0.36**	71.78±0.44
CFPL	**67.11±0.93**	**68.04±1.07**	**80.57±1.26**	**71.91±1.09**	**70.28±1.03**	**75.89±0.32**	93.56±0.10	**79.91±0.48**

Table 2: F1 scores on the StanfordLU dataset under N-way 1-shot and N-way 5-shot settings.

Models	CrossWOZ 1-shot				CrossWOZ 5-shot			
	At	Ho	Re	Avg.	At	Ho	Re	Avg.
TransferM	19.31±0.65	18.24±0.58	18.57±0.77	18.71±0.67	19.79±0.56	18.92±0.86	19.21±0.63	19.31±0.68
MMN	37.16±1.25	35.20±0.83	36.38±2.12	36.25±1.40	39.14±1.48	36.39±0.96	38.21±2.01	37.91±1.48
MPN	38.29±2.07	36.47±1.37	37.42±0.95	37.39±1.46	46.35±3.46	43.26±2.57	49.58±2.33	46.40±2.79
CTLR	47.51±2.18	40.23±1.97	43.78±2.42	43.84±2.19	56.77±3.97	52.34±2.17	71.03±2.94	60.05±3.03
DCKPN	80.62±3.80	68.67±3.33	81.49±3.30	76.93±3.48	83.36±4.62	72.20±3.20	81.40±3.28	78.99±3.70
CFPL	**91.41±0.25**	**77.32±0.57**	**87.09±0.53**	**85.27±0.45**	**90.89±0.49**	**80.45±0.23**	**89.93±0.31**	**87.09±0.35**

Table 3: F1 scores on the CrossWOZ dataset under N-way 1-shot and N-way 5-shot settings.

the performance on the validation domains. For the baseline results on CrossWOZ and JMultiWOZ, we reimplement all the baselines with official codes.

Evaluation Metrics Following previous multi-label few-shot intent detection methods (Zhang et al., 2023), we adopt micro F1 as the metric to evaluate the overall performance. All reported results are the average of 5 different runs.

4.2 Baselines

We evaluate and compare our proposed method with the following strong baselines. **(1) TransferM** is a transfer learning framework (Dai et al., 2007) with a pre-trained language model as the encoder and a multi-layer perceptron as the classifier. It trains on source domains and fine-tunes with support sets from target domains. **(2) Multi-label Prototypical Network (MPN)** represents a modification of the vanilla prototypical network (Snell et al., 2017), which measures the negative Euclidean distance between queries and prototypes, and applies a fixed threshold tuned on dev set for multi-label classification. **(3) Multi-label Matching Network (MMN)** closely resembles MPN but utilizes the Matching Network (Vinyals et al., 2016) to calculate label-instance relevance scores, resulting in classification based on cosine similarity. **(4) CTLR** (Hou et al., 2021) proposes a method for estimating label-instance relevance scores and selecting multiple intent labels using a meta-calibrated threshold, which involves learning universal experience on data-rich domains and adapting thresholds to certain few-shot domains. **(5) DCKPN** (Zhang et al., 2023) constructs a dual class knowledge propagation network that integrates label information and feature structure into graph neural network to guide the intent prediction and employs a multi-label inference method to predict the intent count of each utterance adaptively.

4.3 Main Results

The main results on StanfordLU, CrossWOZ, and JMultiWOZ are shown in Table 2, 3 and 4 respectively. Most baseline results are taken from (Zhang et al., 2023) and the best results are highlighted in bold. We have following observations from the experimental results: (1) CFPL achieves significantly superior average results on three benchmarks across all domains compared to baseline methods, demonstrating the superiority of our method. (2) The performance on 1-shot setting shows more improvements compared to 5-shot setting, which further confirms the efficiency of our method in few-shot tasks. (3) CFPL shows better average 1-shot and 5-shot performance in domains that contain more novel classes (12.9% in Sc domain of StanfordLU and 8.5% in Ho domain of CrossWOZ), indicating that the teacher model introduces more domain features to help recognize unseen classes. (4)

Models	JMultiWOZ 1-shot				JMultiWOZ 5-shot			
	Sh	Ho	Re	Avg.	Sh	Ho	Re	Avg.
TransferM	16.32±1.04	15.85±0.87	14.56±1.52	15.58±1.14	17.11±0.93	16.03±1.08	16.34±1.26	16.49±1.09
MMN	27.21±0.61	23.29±1.54	25.64±1.97	25.38±1.37	32.56±0.74	25.78±1.83	29.30±1.25	29.21±1.27
MPN	30.87±2.84	28.61±3.07	29.67±2.46	29.72±2.79	36.35±2.98	29.14±3.14	34.71±2.77	33.40±2.96
CTLR	27.09±1.47	36.13±2.13	28.30±2.25	30.51±1.95	37.41±1.25	31.95±2.69	31.90±0.57	33.75±1.50
DCKPN	71.61±2.69	57.93±2.22	57.40±2.69	62.31±2.53	70.22±3.38	63.11±1.51	62.36±3.47	65.23±2.79
CFPL	**74.40±0.52**	**63.40±0.48**	**62.91±0.52**	**66.90±0.51**	**76.69±0.66**	**64.13±0.24**	**65.65±0.51**	**68.86±0.47**

Table 4: F1 scores on the JMultiWOZ dataset under N-way 1-shot and N-way 5-shot settings.

Setting	StanfordLU		CrossWOZ		JMultiWOZ	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
CFPL	71.91	79.91	85.27	87.09	66.90	68.86
- LSA	66.07	74.15	78.29	80.54	60.37	62.79
- $Inter$-	65.43	73.52	76.17	78.73	59.85	63.14
- $Intra$-	63.28	69.40	72.14	73.77	56.60	58.13
- $\mathcal{L}_{cir}$	66.87	75.12	77.63	80.38	61.02	64.51

Table 5: Ablation study results. The average F1 scores of all domains are reported.

In terms of the non-English datasets, CFPL outperforms DCKPN 8.2% on CrossWOZ and 4.1% on JMultiWOZ, while eliminating the need of part-of-speech tagging required by CTLR, thereby demonstrating the effectiveness and convenience of our method.

4.4 Ablation Study

To examine the influence of each component, we conduct ablation studies on three datasets, as shown in Table 5. When removing the Intra-domain Prototype Learning (denoted as - $Intra$-), the model performs the worst, indicating the effectiveness of intra-domain feature interaction within each episode for classification. Similarly, when excluding the Inter-domain Prototype Learning (denoted as - $Inter$-), the model exhibits a significant performance decline, indicating that the teacher model introduces unseen domain knowledge, which enhances the ability of recognizing novel classes. When we omit the Label Synset Augmentation (denoted as - LSA), the model performance decreases, indicating that augmented labels help to provide more discriminative representations. When removing the class-specific circle loss (denoted as - $\mathcal{L}_{cir}$), the model performance degrades, indicating the superiority of contrastive loss for multi-label classification.

4.5 Analysis and Discussions

Impact of the Hyperparameter k We conduct sensitivity analysis using different values of k for label augmentation. Figure 3 shows the respective results in different domains and the average results on each dataset. The performance improves significantly as k increases from 0 to 2 in most domains, which indicates the effectiveness of augmented labels. As the value of k continues to increase, the model performance increases very slowly or even decreases after reaching maximum, which implies that excessive label expansion may introduce confusing information or even noise. In addition, an oversized k results in substantial resource consumption. Therefore, we set k = 2 for all the datasets in our experiments.

Parameter Efficiency Due to the computational intensity of Bert-base (110M params) in real industry deployments, we further assess our model performance with the much lighter Bert-tiny (7M params), which is nearly 16 times smaller. Experiments are conducted under different sizes of training set (N-way 1/3/5-shot). The comparison results with the strong baselines in Figure 4 demonstrate that our model still maintains competitive performance when reducing the computational load, reflecting the superiority of our method.

Exploration of LLMs for Multilingual FS-MLID Large language models (LLMs) have achieved significant advancements in numerous few-shot tasks via in-context learning. We conduct a preliminary experiment to explore the performance of LLMs on non-English FS-MLID tasks using gpt-3.5-turbo. Restricted by the input length, we only conduct the N-way 1-shot setting. The experimental results and the in-context learning prompt template that includes task description, demonstration and query are detailed in the Appendix B.

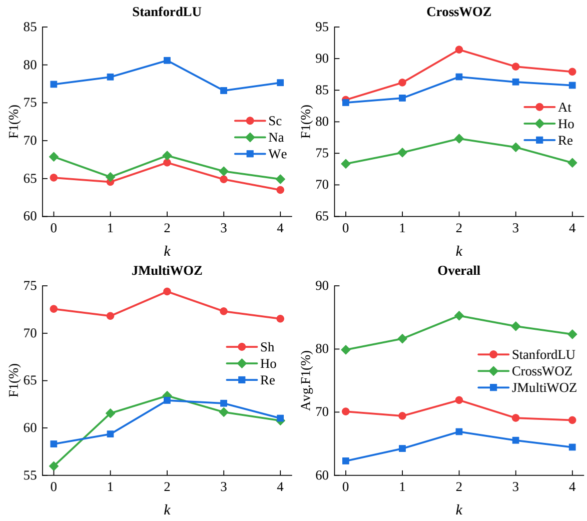


Figure 3: Experimental results of different values of k.

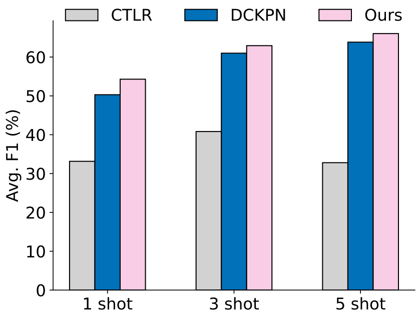


Figure 4: Comparison with CTLR and DCKPN on StanfordLU when using Bert-tiny as the backbone.

5 Conclusion

In this paper, we propose a CFPL method for multi-label few-shot intent detection, which designs a label synset augmentation strategy to enhance the representations of the support data due to the data scarcity issue and proposes to refine the prototypes with knowledge distillation from a universal teacher model. To solve the multilingual intent detection, we fine-tune a cross-lingual teacher model to enable our method to adapt quickly to different languages. To verify our proposed method in detecting intents for multilingual dialogues, we re-annotate two non-English task-oriented dialogue datasets CrossWOZ and JMultiWOZ in multi-label form. Experimental results demonstrate the superiority of our method.

6 Limitations

In this paper, we leverage a multilingual BERT pre-trained on intent corpora as the teacher model. However, we do not explore larger, more powerful generalist language models like LLaMa (Touvron et al., 2023) and Claude (Bai et al., 2022). On the other hand, we hypothesize that related target domain knowledge is compressed in the teacher model, but it may be insufficient for all new domains. Retrieval from related corpora could be a good choice. We leave the exploration of better teacher models and richer target knowledge sources for future study. Additionally, since our method involves some self-attention and cross-attention operations, we plan to speed up the runtime by optimizing attention mechanisms in future work.

7 Ethics Statement

We re-annotate two publicly available non-English task-oriented dialogue datasets, i.e., CrossWOZ and JMultiWOZ, for future multilingual intent detection studies. During the re-annotating process, we make sure that there is no any sensitive information in these datasets, meaning that our work poses no risks to society or individuals.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 62206038, 62106035), Liaoning Binhai Laboratory Project (No. LBLF-2023-01), and Chunhui Project Foundation of the Education Department of China (No. HZKY20220419).

References

Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6548–6557.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *GWC 2016*, pages 50–57.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations*, 19(2):25–35.

Kai-Hsiang Cheng, Szu-Yu Chou, and Yi-Hsuan Yang. 2019. Multi-label few-shot learning for sound event recognition. In *21st IEEE International Workshop on Multimedia Signal Processing*.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *RepL4NLP@ACL 2019*, pages 250–259. Association for Computational Linguistics.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 193–200.

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, NAACL-HLT 2019*, pages 4171–4186.

Mihail Eric, Lakshmi Krishnan, François Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gökhan Tür, and Prem Natarajan. 2023. MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *ACL 2023*, pages 4277–4302.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *NAACL-HLT*, pages 564–569.

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 *NAACL-HLT*, pages 753–757.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393.

Yutai Hou, Yongkui Lai, Yushan Wu, Wanxiang Che, and Ting Liu. 2021. Few-shot learning for multi-label intent detection. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 13036–13044.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *ACL 2022*, pages 2225–2240.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *EMNLP 2023*, pages 12365–12394. Association for Computational Linguistics.

Talaat Khalil, Kornel Kielczewski, Georgios Christos Chouliaras, Amina Keldibek, and Maarten Versteegh. 2019. Cross-lingual intent classification in a low resource industrial setting. In *EMNLP-IJCNLP 2019*, pages 6418–6423.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1079–1087.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, and Xianchao Zhang. 2021. An explicit-joint and supervised-contrastive learning framework for few-shot intent classification and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1945–1955.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, pages 2773–2782.

Atsumoto Ohashi, Ryu Hirai, Shinya Iizuka, and Ryuichiro Higashinaka. 2024. Jmultiwoz: A large-scale japanese multi-domain task-oriented dialogue dataset. In *LREC/COLING 2024*, pages 9554–9567. ELRA and ICCL.

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 *EMNLP-IJCNLP*, pages 2078–2087.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021. GL-GIN: fast and accurate non-autoregressive model for joint multiple intent detection and slot filling. In *ACL/IJCNLP*, pages 178–188.

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI 2020*, pages 8689–8696. AAAI Press.

Barbara Rychalska, Helena T. Glabska, and Anna Wróblewska. 2018. Multi-intent hierarchical natural language understanding for chatbots. In *International Conference on Social Networks Analysis, Management and Security, SNAMS*, pages 256–259.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. 2021. Enhancing the generalization for intent classification and out-of-domain detection in SLU. In *ACL 2021*, pages 2443–2453.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 4077–4087.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6397–6406.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 3630–3638.

Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Pawel Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *EMNLP 2022*, pages 7544–7559.

Ivan Vulic, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Pawel Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *EMNLP 2022*, pages 7544–7559.

Zichen Wang, Bo Yang, Haonan Yue, and Zhenghao Ma. 2024. Fine-grained prototypes distillation for few-shot object detection. In *AAAI 2024*, pages 5859–5866.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP 2019*, pages 6381–6387.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding. In *EMNLP*, pages 4884–4896.

Yangjun Wu, Han Wang, Dongxiang Zhang, Gang Chen, and Hao Zhang. 2022. Incorporating instructional prompts into a unified generative framework for joint multiple intent detection and slot filling. In *COLING 2022*, pages 7203–7208.

Shuo Yang, Lu Liu, and Min Xu. 2021. Free lunch for few-shot learning: Distribution calibration. In *ICLR 2021*. OpenReview.net.

Feng Zhang, Wei Chen, Fei Ding, and Tengjiao Wang. 2023. Dual class knowledge propagation network for multi-label few-shot intent detection. In *ACL 2023*, pages 8605–8618.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022. Prompt-based meta-learning for few-shot text classification. In *EMNLP 2022*, pages 1342–1357.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Trans. Assoc. Comput. Linguistics*, 8:281–295.

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 *WSDM*, pages 1022–1031.

A Dataset Re-annotation Details

We re-annotate the two public multi-domain task-oriented dialogue datasets, i.e., CrossWOZ (Zhu et al., 2020) and JMultiWOZ (Ohashi et al., 2024), into multi-label form. Specifically, in order to align with the setting of the multi-label intent detection task, we have expanded the original task-oriented dialogue dataset which includes the basic user intent in the form of multiple labels. The utterance in both non-English multi-domain task-oriented dialogue datasets is originally labeled with one simple user intent, but it actually has more than one intent labels. The entire process of re-annotation consists of three stages: first, we follow the existing multi-label intent datasets (Qin et al., 2020; Rastogi et al., 2020) to determine multiple more precise intent labels; then, we re-annotate the two public multi-domain task-oriented dialogue datasets; finally, we carefully review the re-annotated data.

In our re-annotation process, we first identify the domains of the FS-MLID task in two non-English datasets. The principle is to make the domains of the two datasets as similar as possible to evaluate the generalizability of our method across different languages. For the CrossWOZ dataset, the three re-annotated domains are At (Attraction), Ho (Hotel), and Re (Restaurant). For the JMultiWOZ dataset, the three re-annotated domains are Sh (Shopping), Ho (Hotel), and Re (Restaurant). Due to the constraints imposed by the length of user utterances, the intents expressed by users in the two non-English datasets are not as diverse as those in the StanfordLU. This lead to fewer label classes being identified in our study compared to the StanfordLU. For the two non-English datasets, we engage three native speakers of each dataset's language as annotators to reannotate the sentences with multi-label annotations. This process spans five days. For the CrossWOZ dataset, a total of 8697 user utterances are reannotated, and for the JMultiWOZ dataset, a total of 8076 utterances are reannotated. Finally, the reannotated results undergo a manual review process by two proficient speakers of each language to ensure accuracy.

B LLMs for Multilingual FS-MLID Tasks

LLMs Prompt Template in N-way 1-shot Setting Given a target user intent list from task-oriented dialogue, an user utterance, please identify all intents behind user utterances. Note that the setting of this task conforms to the N-way 1-shot setting, which includes two stages: meta-training and meta-testing. In the meta-training phase, there are 200 few-shot episodes from source domain. In the meta-testing phase, there are 50 few-shot episodes from target domain. Each episode contains a support set and a query set, and the query set size is 32.

Target user intent list of source domain:
<1>: <User intent 1>
<2>: <User intent 2>
......
<N_s>: <User intent N_s>

Episodes in meta-training phase from source domain (Each episode contains support set and query set):

<Episode 1>:{"support":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......
<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......
......},
"query":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......
<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......
......
<Utterance 32>: <User intent 1 of Utterance 32>; <User intent 2 of Utterance 32>; ......} };
<Episode 2>:{"support":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......
<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......
......},
"query":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......
<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......
......
<Utterance 32>: <User intent 1 of Utterance 32>; <User intent 2 of Utterance 32>; ......} };
......
<Episode 200>:{"support":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......
<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......
......},
"query":{
<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......

<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......

......

<Utterance 32>: <User intent 1 of Utterance 32>; <User intent 2 of Utterance 32>; ......} }

Target user intent list of target domain:

<1>: <User intent 1>

<2>: <User intent 2>

......

<N_t>: <User intent N_t>

Episodes in meta-testing phase from target domain (Each episode contains support set and query set):

<Episode 1>:{"support":{

<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......

<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......

......},

"query":{

<Utterance 1>: <User intent ID>; <User intent ID>; ......

<Utterance 2>: <User intent ID>; <User intent ID>; ......

......

<Utterance 32>: <User intent ID>; <User intent ID>; ......} };

<Episode 2>:{"support":{

<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......

<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......

......},

"query":{

<Utterance 1>: <User intent ID>; <User intent ID>; ......

<Utterance 2>: <User intent ID>; <User intent ID>; ......

......

<Utterance 32>: <User intent ID>; <User intent ID>; ......} };

......

<Episode 50>:{"support":{

<Utterance 1>: <User intent 1 of Utterance 1>; <User intent 2 of Utterance 1>; ......

<Utterance 2>: <User intent 1 of Utterance 2>; <User intent 2 of Utterance 2>; ......

......},

"query":{

<Utterance 1>: <User intent ID>; <User intent ID>; ......

<Utterance 2>: <User intent ID>; <User intent ID>; ......

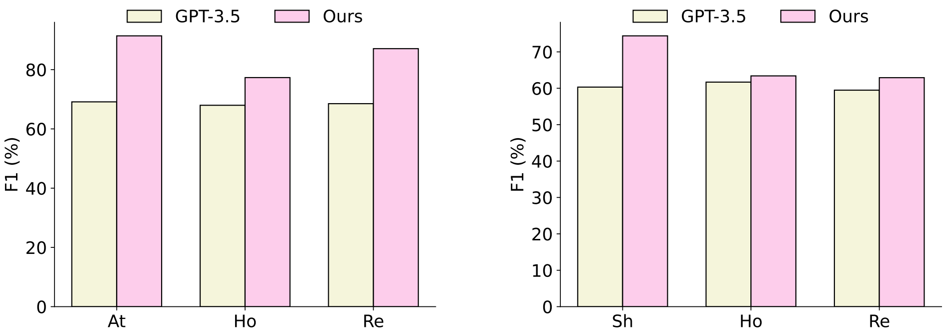
......

<Utterance 32>: <User intent ID>; <User intent ID>; ......} }

Experimental Results and Analysis We conduct experiments using gpt-3.5-turbo [2] on two non-English datasets: CrossWOZ and JMultiWOZ. We input the data and instructions into gpt-3.5-turbo (GPT-3.5) according to the N-way 1-shot setting, where N is the number of classes in each domain. To simulate the meta-learning paradigm of few-shot learning, we divide the entire process into the meta-training and the meta-testing phase. All the few-shot episode construction is consistent with our experiment. For the meta-training phase, we construct 200 few-shot episodes for each source domain and 50 few-shot episodes for each target domain. And the size of query set is 32. We input GPT-3.5 with an user intent list of each source or target domain and the task descriptions, and the requested response is the intent ID of each query user utterance.

The experimental results on CrossWOZ and JMultiWOZ are shown in Figure 5. From the results, it can be observed that CFPL performs much better than GPT-3.5, which indicates the superiority of our method. During the experiment, we observe that some of the responses provided by GPT-3.5 are blank, which has a substantial impact on accuracy. Additionally, we observe that the capability of GPT-3.5 for predicting the number of multiple labels still requires further improvement. It is very intuitive to observe that the performance of GPT-3.5 on FS-MLID task is more related to languages rather than specific domains. Specifically, the average F1 scores of GPT-3.5 are 68.54% on CrossWOZ and 62.48% on JMultiWOZ. However, for different domains within the same dataset, the maximum and minimum F1 scores only differ by 1.2% on the CrossWOZ dataset and 2.2% on the JMultiWOZ dataset. Inspired by (Huang et al., 2023) and combined with our experimental results, we analyze the reason is that LLMs do not have equal capability of handling all the languages, leading to imbalanced performance across different languages. This further underscores the importance of dealing with intent detection from a multilingual perspective.

[2] `https://platform.openai.com/docs/models/gpt-3-5`



(a) Results on CrossWOZ

(b) Results on JMultiWOZ

Figure 5: Results of LLMs for multilingual FS-MLID tasks.

Datasets	CTLR (train)	CTLR (test)	DCKPN (train)	DCKPN (test)	CFPL (train)	CFPL (test)	Cross-Lingual (CFPL)
StanfordLU (1-shot)	25	10	24	6	86	26	/
StanfordLU (5-shot)	32	13	28	5	87	27	/
CrossWOZ (1-shot)	44	11	36	7	128	27	350
CrossWOZ (5-shot)	56	10	38	8	130	28	350

Table 6: Total training and inference time (wall clock time in second).

C Error Rate Analysis

The corresponding class label sets for schedule, navigate and weather domains are as follows: [‘request_location’, ‘inform’,‘query’, ‘confirm’, ‘appreciate’, ‘command_appointment’, ‘remind’, ‘request_information’, ‘list_schedule’, ‘request_time’, ‘request_party’, ‘request_agenda’, ‘schedule’, ‘request_date’], [‘request_poi’, ‘inform’, ‘query’, ‘confirm’, ‘appreciate’, ‘request_address’, ‘request_route’, ‘request_traffic’, ‘show_in_screen’, ‘navigate’], and [‘request_low_temperature’, ‘request_time’, ‘appreciate’, ‘request_temperature’, ‘request_weather’, ‘inform’, ‘request_high_temperature’, ‘query’]. These three domains all contain semantically highly similar class labels, such as ‘request_time’ and ‘request_date’ in the schedule domain. Moreover, there are also class labels with highly similar structures and meanings, such as ‘list_schedule’ and ‘schedule’ in the schedule domain, ‘request_poi’ and ‘request_address’ in the navigate domain. Our error rate for similar categories is significantly lower than that of the strong baseline DCKPN. Taking the schedule domain as example, for our method CFPL, the probability of the data belonging to ‘request_time’ that are misclassified into ‘request_date’ is 6.7%, whereas DCKPN has a more higher error rate, which is 9.1%. The class label with the highest error rate in our method is ‘query’ in all three domains. Its error rate is 16.2% in the schedule domain, 12.1% in the navigate domain, and 8.3% in the weather domain. The basic reason is that users express questions in diverse forms, so the method has weak ability to classify ‘query’. We also observe that in the scheduling domain, the error rate of ‘schedule’ is very high at 14.9%, the reason is that the label set contains several labels with similar meanings (e.g.‘request_agenda’, ‘list_schedule’, etc), which interferes with the classification of the data belonging to ‘schedule’.

D Runtime Analysis

CFPL involves several non-trivial steps, which impact the total wall clock time. The total training and inference time is summarized in Table 6. It can be seen that our method requires more time compared to the baselines, particularly in the training phase. The increase in time can be attributed to the multiple parts of CFPL. Although CFPL requires more time, it provides more accurate and robust results, especially in the cross-lingual scenarios.

Since our method involves some self-attention and cross-attention operations during inference, which takes up a lot of time. In future work, we plan to speed up inference time by optimizing attention mechanisms. This will involve reducing both the computational and memory overheads associated with attention calculations, as well as minimizing the memory access costs related to IO operations. By focusing on these areas, our method could achieve significant speed improvements.