A Two-Stage Prediction-Aware Contrastive Learning Framework for Multi-Intent NLU

Guanhua Chen, Yutong Yao, Derek F. Wong[†], Lidia S. Chao

NLP2CT Lab, Department of Computer and Information Science, University of Macau

{nlp2ct.guanhua, nlp2ct.yutong}@gmail.com

{derekfw, lidiasc}@um.edu.mo

Abstract

Multi-intent natural language understanding (NLU) presents a formidable challenge due to the model confusion arising from multiple intents within a single utterance. While previous works train the model contrastively to increase the margin between different multi-intent labels, they are less suited to the nuances of multi-intent NLU. They ignore the rich information between the shared intents, which is beneficial to constructing a better embedding space, especially in low-data scenarios. We introduce a two-stage Prediction-Aware Contrastive Learning (PACL) framework for multi-intent NLU to harness this valuable knowledge. Our approach capitalizes on shared intent information by integrating word-level pre-training and prediction-aware contrastive fine-tuning. We construct a pre-training dataset using a word-level data augmentation strategy. Subsequently, our framework dynamically assigns roles to instances during contrastive fine-tuning while introducing a prediction-aware contrastive loss to maximize the impact of contrastive learning. We present experimental results and empirical analysis conducted on three widely used datasets, demonstrating that our method surpasses the performance of three prominent baselines on both low-data and full-data scenarios.

1. Introduction

Multi-intent Natural Language Understanding (NLU) models are fundamental building blocks within task-oriented dialogue systems (Qin et al., 2019; Gangadharaiah and Narayanaswamy, 2019). These systems encompass multi-intent detection (mID) and slot-filling (SF) tasks. However, effectively capturing multiple intents within short utterances presents a formidable challenge, primarily attributed to limited labeled data and the vast spectrum of spoken expressions. Consequently, a plethora of advanced models have emerged to refine the granularity of dialogue content and investigate the relationships among different intents (Qin et al., 2020; Song et al., 2022). Furthermore, recent researches (Qin et al., 2021; Chen et al., 2022a; Cai et al., 2022a; Wu et al., 2022) also proved that the extensive linguistic knowledge embedded within these pre-trained models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), is effective to facilitate the comprehension of multiple intents within brief utterances.

However, these models often face a trade-off between inference speed and the incorporation of additional modules aimed at capturing more knowledge in the training data. This knowledge assumes a pivotal role in facilitating the model's acquisition of a more distinguishable semantic representation space, thereby yielding substantial benefits for downstream tasks. Consequently, several training strategies have emerged to maximize the utility of existing data without affecting the inference speed, such as contrastive learning (Liu et al., 2021; Basu et al., 2022) or curriculum learning (Chen et al., 2023; Zhou et al., 2020; Zhan et al., 2021). Leveraging these methods empowers the model to construct an improved embedding space while utilizing the same volume of data.

Regrettably, existing contrastive learning (CL) methods on multi-intent NLU typically assign fixed roles, either positive or negative, to samples. This potentially disregards the valuable knowledge embedded in the relationships between the shared intents, a factor that has been demonstrated to enhance multi-intent detection (Xu and Sarikaya, 2013). Moreover, fine-tuning the model through equal-weight contrastive learning cannot fully learn the knowledge in the training data.

To tackle the above issues, we propose a two-stage Prediction-Aware Contrastive Learning (PACL) framework. PACL is meticulously designed to effectively harness knowledge emanating from shared intents through two-stage training: word-level pre-taining and prediction-aware contrastive fine-tuning. Building on Xu and Sarikaya (2013) proof of words commonly used to express an intent frequently appear across various instances, we further observed that these words exhibit a distinctive emphasis within the Part-of-Speech (POS) distribution. Hence, we introduce a word-level data augmentation strategy to construct a dataset for self-supervised pre-training. It enables the model to learn the associations between meaningful words

---

[†] Corresponding authors.

and corresponding intents accurately. For fine-tuning stage, we devise an innovative prediction-aware contrastive learning framework. It facilitates automatic role-switching for each sample, allowing it to dynamically alternate between positive and negative roles while adjusting its influence based on the model's confidence. Additionally, we design an intent-slot attention mechanism to establish a strong connection between the mID and SF tasks for contrastive learning, which incentivized to more effectively harness the knowledge derived from shared intents, culminating in an embedding space characterized by heightened distinguishability.

We evaluate our PACL framework on three prominent multi-intent datasets: MixATIS (Qin et al., 2020), MixSNIPS (Qin et al., 2020), and StanfordLU (Hou et al., 2021). In addition, we employ three robust baselines for comparison: RoBERTa (Liu et al., 2019), TFMN (Chen et al., 2022b), and SLIM (Cai et al., 2022a). Our experimental results demonstrate that PACL yields substantial enhancements in model performance while accelerating convergence speed. Our empirical analyses further reaffirm the indispensability of each component within our framework.

2. Related work

2.1. Multi-Intent NLU

Compared to single-intent detection (Zhang et al., 2019; Wu et al., 2020; Cheng et al., 2021), multiple intents appear more common in real-time scenarios. Early research (Kim et al., 2017; Gangadharaiah and Narayanaswamy, 2019) attempts to apply traditional CNN or RNN-based methods. Qin et al. (2020) proposed An Adaptive Graph-Interactive Framework (AGIF), and further expanded this technique to a non-autoregressive model (Qin et al., 2021). Cai et al. (2022b) proposed Explicit Slot-Intent Mapping with Bert (SLIM) which transferred JointBERT from single intent to multi-intent task while solving the shared-intent problem. Considering the number of intents is important in the multi-intent NLU task, Chen et al. (2022b) designed a novel threshold-free framework to predict the number of intentions in the utterance before predicting the specific intents.

2.2. Contrastive Learning

Contrastive Learning (CL) has been widely used on NLU tasks, because of its data scarcity and diverse expression. The recent works show the effect of CL on the NLU task (Gunel et al., 2020; Hou et al., 2021; Yehudai et al., 2023). Specifically, for the multi-intent NLU task, Vulić et al. (2022) devised a strategy to transform a general sentence-encoder into a task-specific one on multi-intent
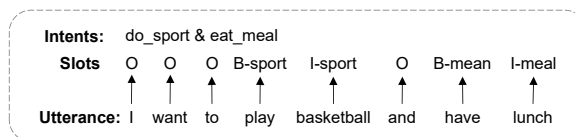

Figure 1: An example of multi-intent NLU task.

data through contrastive learning. Tu et al. (2023) proposed a novel bidirectional joint model trained using supervised CL and self-distillation, effectively utilizing intent and slot features to complement each other.

3. Proposed Method

3.1. Problem Formulation

As illustrated in Figure 1, given an input utterance $x = (x_1, x_2, ..., x_n)$ with n tokens, the multi-intent NLU model entails the simultaneous prediction of both multi-label intents for the utterance and the slot-filling for each word. Multi-label intent signifies the presence of more than one distinct intent within the set of possible intents.

The effectiveness of a joint training strategy for NLU task has been proved by Chen et al. (2019). The joint objective functions can be mathematically formulated as follows:

$$\mathcal{L}_{joint} = \mathcal{L}_{ID} + \mathcal{L}_{SF} \tag{1}$$

where $\mathcal{L}_{ID}$ and $\mathcal{L}_{SF}$ are the cross entropy loss of intent detection and slot filling tasks.

Building upon this foundation, we introduce our two-stage prediction-aware contrastive learning (PACL) framework as shown in Figure 2.

3.2. Word-level Pre-training

As multi-intent samples tend to be particularly challenging to distinguish within the embedding space, our initial step involves word-level pre-training to bolster the model's adaptability to the specific domain. Based on Cai et al. (2022a), which released the correspondence between each token and sub-intent in the MixATIS (Qin et al., 2020) and MixSNIPS (Qin et al., 2020) datasets, we conducted an analysis focusing on the distribution of Part of Speech (POS) categories associated with tokens linked to distinct intents. Notably, POS categories like "NN", "NNS", and "JJ" accounted for a significant proportion of words connected to intent labels, reaching as high as 87.76% in MixATIS and 78.97% in MixSNIPS. This unveiled a strong correlation between specific POS categories and intents. We found that the intents are strongly correlated with some words with specific POS.
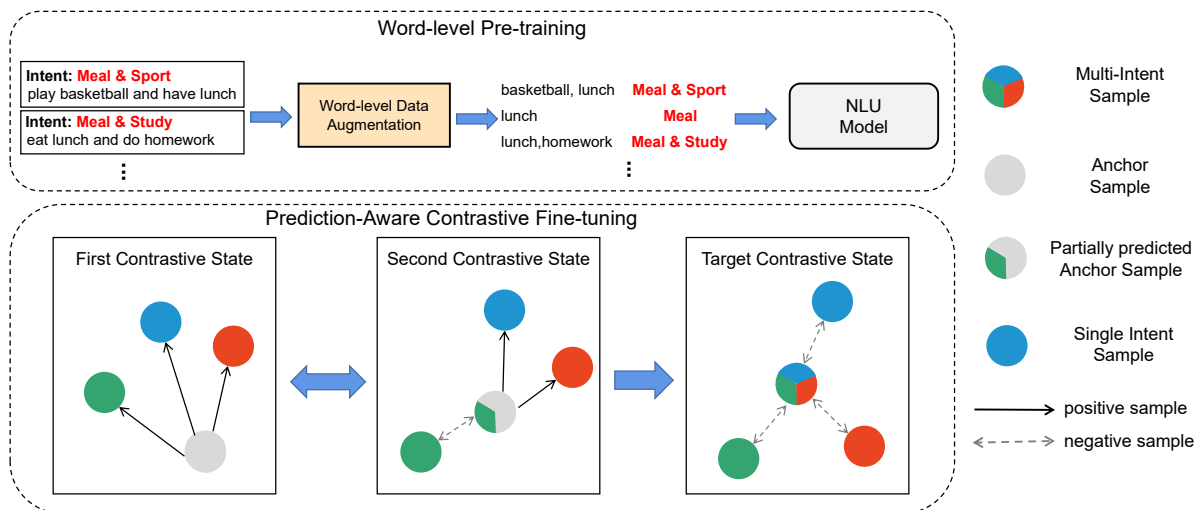
Figure 2: The overview of our framework. Different shapes indicate completely different samples and different shades of color indicate samples with shared intent.

Thus, we split the original utterance-level dataset into word-level, concentrating on those aligned with the specified POS categories to construct a word-level multi-intent dataset. Firstly, we detect whether a word with the specific POS recurs across multiple utterances. Subsequently, we refined the intent associated with this word by extracting the shared intent from those recurring utterances. For instance, if a term "lunch" occurs in utterances linked to both "$Meal\ \&\ Sport$" and "$Meal\ \&\ Study$" intents, we identify it as belonging to "$Meal$". For words that remained connected to multiple intents, we concatenate them with words specifically indicating the associated intent. Since this is word-level pre-training, it is acceptable to have two words associated with the same intent in an input even after concatenation.

Due to the absence of sentence structure information in the concatenation of words, our pre-training strategy exclusively focuses on the intent detection task. This phase is dedicated to enabling the model to acquire an understanding of the relationships between individual words and various intents, and facilitates the model's ability to more effectively capture the presence of multiple intents within an utterance. The final pre-training loss can be formulated as:

$$\mathcal{L}_{PT} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}_{CL} \qquad (2)$$

where $\mathcal{L}_{ID}$ is the cross-entropy loss of intent prediction and $\mathcal{L}_{CL}$ is the traditional contrastive loss of the intent logits. $\{\lambda_1, \lambda_2\}$ are the hyper-parameters to balance each loss.

3.3. Prediction-Aware Contrastive Fine-tuning

Upon the word-level pre-trained model, we introduce an innovative prediction-aware contrastive loss to fine-tune the model on the original utterance-level dataset. In this fine-tuning process, each instance, which shares common intents with other samples, dynamically alternates its role (either positive or negative).

As depicted in Figure 2, given an anchor sample, samples with incorrectly predicted intents are designated as positive samples, while those with correctly predicted intents serve as negative ones. To illustrate, consider a sample with the intent "$atis\_capacity\#atis\_city$", once the model successfully predicts part of the intent, like "$atis\_capacity$", our model categorizes samples with "$atis\_capacity\#atis\_city$" and "$atis\_capacity$" as the positive candidates. Simultaneously, any samples with different intents, including "$atis\_capacity$", are treated as negative. Because the anchor sample has already learned the knowledge of the relationship between the anchor sample and intent "$atis\_capacity$".

Noteworthy, those instance pairs with identical or completely different multi-intent labels will be regarded as mutually positive or negative pairs during the entire training. This prediction-aware contrastive learning empowers the model to gain a substantial understanding of the relationships between shared intents while expediting the clustering of samples sharing common intents.

In greater detail, each time we construct the mini-batch, we sample a maximum of K positive samples for each anchor sample. Regarding the negative samples, we randomly sampled them from the mini-batch. In situations involving a multi-intent

label to which only the anchor sample belongs, we forward propagate this sample twice, following SimCSE (Gao et al., 2021), to secure at least one positive sample for each anchor. The original CL objective function can be defined as:

$$\mathcal{L}_{CL}^i = \sum_{p \in P(i)} \log \frac{f(h_i, h_{i,p})}{\sum_{k \in H(i)} f(h_i, h_{i,k})} \quad (3)$$

where $f(h_1, h_2) = \exp(cos(h_1, h_2)/\tau)$. $P(i)$ is the set of positive samples, while $H(i)$ denotes the set of all positive and negative samples for the i-th anchor sample. h_i is the representation vector for contrastive learning. The τ is the temperature value in CL loss. In our experiments, we sample at most K positive or negative samples.

Nevertheless, contend that positive and negative samples with varying predicted probabilities should exert distinct influences on contrastive learning. We leverage the predicted probabilities to calibrate the impact of each sample within our prediction-aware contrastive loss function. In practice, the higher the average probability of the intents associated with the anchor sample, which matches that of the negative sample, the more challenging it becomes to distinguish the negative sample within the embedding space. Hence, we normalize the probability as the weight of negative samples as represented by the following equation:

$$w_{i,n} = \text{Softmax}(Mean(\{p_i | y_i^{neg}\})) \quad (4)$$

where $p_{i,j}$ is the probability that the i-th sample in the batch belongs to the j-th intent.

With respect to the positive samples, a higher probability of a shared intent between the positive sample and the anchor sample signifies the model's enhanced capacity for precise classification. Consequently, it implies that the model can afford to allocate reduced attention to these samples. Correspondingly, we assign weights to the positive samples as follows:

$$w_{i,p} = \text{Softmax}(\alpha_i \prod \{p_i | y_i^{pos}\}),$$
$$\alpha_i = \frac{\text{Count}(y_i^{pos})}{\text{Count}(y_i)} \quad (5)$$

where the $\text{Count}(\cdot)$ means the number of elements that satisfy the condition, while y_i refers to the predicted intent, and y_i^{pos} represents the correctly predicted intent. Combining the equation 4 and 5, the prediction-aware contrastive loss can be rewritten as:

$$\mathcal{L}_{PACL}^i = \sum_{p \in Pos(i)} \log \frac{w_{i,p} f(h_i, h_{i,p})}{\sum_{k \in H(i)} w_{i,k} f(h_i, h_{i,k})} \quad (6)$$

Finally, in order to enhance the connection between multi-intent detection and slot filling tasks, we add a multi-head attention layer to capture the relationship between all the slots and intent representation. The final representation for CL can be generated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V, \quad (7)$$
$$h_{intent} = \text{Attention}(h_{cls}, H_{slot}, H_{slot})$$

$$h'_{intent} = (W[h_{intent}, h_{cls}] + b) \quad (8)$$

where $\text{Attention}(\cdot)$ is the attention mechanism as described by Vaswani et al. (2017). $H_{slot} \in \mathbb{R}^{n \times d}$ is the set of all the word representations of the given utterance. W is the trainable parameters, and $h'_{intent} \in \mathbb{R}^d$ is the final representation. We feed the h'_{intent} to the same intent classifier in equation 1 to train the additional layer in equation 7 and 8 only, which has few parameters and will not affect the training speed too much. This allows the latent space used for the CL to be more adapted to the intent classifier in equation 1.

Overall, the final loss of the fine-tuning process can be formulated as:

$$\mathcal{L}_{FT} = \lambda_3 \mathcal{L}_{ID} + \lambda_4 \mathcal{L}_{SF} + \lambda_5 \mathcal{L}_{PACL} \quad (9)$$

where $\{\lambda_3, \lambda_4, \lambda_5\}$ are the hyper-parameters to balance the impact of each loss.

4. Experiments

4.1. Dataset and Evaluation

We evaluate our method on three widely used multi-intent datasets, **MixATIS** (Qin et al., 2020), **MixSNIPS** (Qin et al., 2020), and **StanfordLU** (Hou et al., 2021). MixATIS contains 13161/756/828 utterances in train/validation/test set. MixSNIPS contains 39776/2198/2199 utterances in train/validation/test set. As for StanfordLU, it consists of 6428/790/820 train/validation/test data. We sampled 10% of each intent for our low-data scenarios. For evaluation metrics, we use F1 score, intent accuracy, and intent-slot overall accuracy.

4.2. Baselines

We compare our proposed approach with previous state-of-the-art models (E et al., 2019; Qin et al., 2019, 2020, 2021), whose results are taken from previous work directly. Besides, we also reproduce three strong baselines: **RoBERTa**[1] (Liu et al., 2019), **TFMN**(Chen et al., 2022b), and **SLIM**[2] (Cai

---
[1]https://huggingface.co/roberta-base
[2]https://github.com/TRUMANCFY/SLIM

Model	MixATIS			MixSNIPS		
	IC Acc	SF F1	Overall Acc	IC Acc	SF F1	Overall Acc
AGIF(Qin et al., 2020)	75.8	88.1	44.5	96.5	94.5	76.4
GL-GIN (Qin et al., 2021)	76.3	88.3	43.5	95.6	94.9	75.4
GISCo (Song et al., 2022)	-	-	48.2	-	-	75.9
SDJN+BERT (Chen et al., 2022a)	78.0	87.5	46.3	96.7	95.4	79.3
Low-Data						
RoBERTa[†] (Liu et al., 2019)	46.7	**84.1**	24.9	94.0	92.3	67.5
RoBERTa (PACL)	**66.2**	84.0	**36.8**	**94.8**	**92.9**	**69.6**
TFMN[†] (Chen et al., 2022b)	46.4	71.1	15.7	95.2	90.9	65.2
TFMN (PACL)	**58.3**	**72.7**	**18.1**	**95.5**	**91.9**	**67.8**
SLIM[†] (Cai et al., 2022a)	52.8	72.4	16.1	93.9	93.4	73.8
SLIM (PACL)	**60.8**	**75.8**	**23.3**	**95.2**	**94.4**	**75.4**
Full-Data						
RoBERTa[†] (Liu et al., 2019)	77.5	85.5	47.7	**96.5**	95.6	80.9
RoBERTa (PACL)	**79.1**	**86.0**	**48.9**	**96.5**	**96.2**	**83.4**
TFMN[†] (Chen et al., 2022b)	81.5	**86.9**	48.6	97.1	96.1	83.4
TFMN (PACL)	**82.9**	86.7	**49.4**	**97.4**	**96.3**	**83.6**
SLIM[†] (Cai et al., 2022a)	78.9	87.1	46.4	96.8	96.3	83.7
SLIM (PACL)	**81.9**	**87.3**	**50.4**	**96.9**	**96.8**	**85.1**

Table 1: The results (%) on the test set of two datasets. [†] means we reproduce this framework as our baselines. **Bold** numbers indicate the better result for each baseline, meanwhile, underlined numbers show the best performance in the column.

et al., 2022a), to verify the effect of our framework. More details about our implementations will be explained in Appendix.

4.3. Implementation Detail

We use the BERT-Base[3] model for SLIM (Cai et al., 2022a) and TFMN (Chen et al., 2022b). Besides, we also employ the RoBERTa-Base[4] (Liu et al., 2019) model with 2 feed-forward classifier released by hugging face. The max sequence length and batch size are 50 and 64 for both pre-training and fine-tuning. The sampling number K in our experiments is set to 5. We pre-train the model for $\{1, 3, 5\}$ epochs and fine-tune it for 10 epochs. For Roberta and TFMN models, we set the dropout rate and learning rate to 0.1 and 2e-5. For the SLIM model, we set the dropout rate and learning rate to 0.2 and 5e-5. For more details about training, given an input utterance with n tokens, $x = \{x_1, x_2, ..., x_n\}$, and its encoded representation, $h = \{h_{cls}, h_1, h_2, ..., h_n, h_{sep}\}$, where the representation of the [CLS] token is utilized for multi-intent detection, and the representations of the other tokens are utilized for slot filling, separately.

---

[3]https://huggingface.co/bert-base-uncased
[4]https://huggingface.co/roberta-base

Model	StanfordLU		
	IC Acc	SF F1	Overall Acc
Low-Data			
RoBERTa[†] (Liu et al., 2019)	21.3	**29.1**	21.2
RoBERTa (PACL)	**32.4**	28.8	**31.9**
TFMN[†] (Chen et al., 2022b)	41.7	30.0	38.8
TFMN (PACL)	**46.7**	**31.8**	**40.2**
Full-Data			
RoBERTa[†] (Liu et al., 2019)	88.0	92.3	82.9
RoBERTa (PACL)	**89.0**	**92.5**	**84.1**
TFMN[†] (Chen et al., 2022b)	88.0	**93.0**	83.6
TFMN (PACL)	**89.1**	92.9	**84.3**

Table 2: The results (%) on the test set of StanfordLU dataset. [†] means we reproduce this framework as our baselines. **Bold** numbers indicate the better result for each baseline.

4.4. Main Results

The primary experimental findings are summarized in Table 1 and Table 2. Our proposed framework consistently outperforms RoBERTa (Liu et al., 2019), TFMN (Chen et al., 2022b), and SLIM (Cai et al., 2022a) on both low-data and full-data scenarios. Notably, we conducted experiments solely with the RoBERTa and TFMN models on the StanfordLU dataset, as this dataset lacks matching data for intent and individual tokens. Intuitively, the improvements achieved by the PACL framework are more substantial in low-data scenarios compared to high-data scenarios, across all three evaluation

Method			MixATIS					
			Low-data			Full-data		
PT	RS	PA	IC Acc	SF F1	Overall Acc	IC Acc	SF F1	Overall Acc
Baseline			52.8	72.4	16.1	78.9	87.1	46.4
-	-	-	49.6	71.6	16.1	79.8	86.9	48.1
√	-	-	57.4	73.3	19.3	79.6	86.1	47.8
-	√	-	48.3	77.0	19.1	80.3	86.9	47.5
-	-	√	50.1	71.5	15.6	80.1	**87.6**	48.1
√	√	-	54.7	75.4	21.7	80.5	86.9	48.9
√	-	√	54.5	75.4	21.3	79.9	86.4	49.3
-	√	√	58.8	71.2	17.6	81.6	87.5	50.1
√	√	√	**60.8**	**75.8**	**23.3**	**81.9**	87.3	**50.4**

Table 3: The ablation study results(%). None of the "PT", "RS", or "PA" components are used meaning the traditional CL method. More explanations are in Section 5.1.

metrics. For instance, on the MixATIS dataset, our PACL method demonstrates a remarkable 11.9% increase in intent detection accuracy and a 7.2% boost in overall accuracy in the low-data scenario, whereas it only achieves an improvement of up to 2.7% in intent detection accuracy and 3.7% in overall accuracy in the high-data scenario.

Furthermore, we anticipate a modest enhancement in slot-filling F1 scores, as the primary objective of combining intent and slot representations is to primarily enhance overall accuracy. Consequently, we observe instances where the overall accuracy improved, while the individual scores for intent detection and slot filling exhibited less significant improvements. It is pertinent to mention that the improvement in intent classification accuracy is relatively lower on the MixSNIPS dataset compared to the other two datasets. This divergence may be attributed to the fact that MixATIS and StanfordLU comprise a significantly larger number of multi-intent labels, with 17 intents and 24 intents, in contrast to MixSNIPS, which contains only 6 intents. This discrepancy increases the proportion of overlapping intents, which, in turn, results in our PACL framework performing more effectively on the MixATIS and StanfordLU datasets.

5. Analysis

5.1. Ablation Study

For a better explanation, we divide the fine-tuning stage into two components in this section: dynamic role selection for each sample and the integration of prediction-aware contrastive loss. We conducted ablation experiments in both low-data and full-data scenarios to assess the efficacy of each component, as summarized in Table 3. In this context, "PT" refers to the word-level pre-training process, "RS" signifies the dynamic selection of roles (positive or negative) for each sample during training, and "PA" indicates the prediction-aware contrastive loss. Our findings unequivocally establish the necessity of each component. Specifically, "RS" and "PA" can both contribute to enhancing the performance across all three metrics, while combining "PA" and "RS" obtains further improvement on overall performance. This indicates that these two components can benefit each other. Regarding "PT", it facilitates the model in acquiring domain-specific precise knowledge at word-level, consequently bolstering the effectiveness of the subsequent prediction-aware contrastive fine-tuning. However, "PT" gets less improvement on full data than on low-data, due to the fact that the greater amount of data in the full-data scenario allows the model to learn the knowledge between shared intents better, which diminishes the effect of "PT".

5.2. Different Proportions of Training Data

To further explore the effectiveness of our method on different data volumes, we conducted experiments on MixATIS dataset with {20%,30%,40%,50%} data, and the results compared with baselines are shown in Table 4. Our method has an average 3% improvement in intent accuracy as well as an average 3.28% improvement in overall accuracy for data on different proportions. Intuitively, the enhancement of our method slightly decreases as the proportion of data increases. It is because more training data will enable the model to recognize multiple intents more clearly, which in turn diminish the effect of

Percentage	MixATIS		
	IC Acc	SF F1	Overall Acc
10%	52.8	72.4	16.1
	60.8	**75.8**	**23.3**
20%	71.4	82.1	36.9
	72.6	**84.2**	**40.2**
30%	74.8	84.7	42.3
	75.9	**86.2**	**44.0**
40%	75.7	85.9	43.3
	77.9	**86.9**	**45.7**
50%	75.8	86.1	44.6
	78.0	**86.9**	**46.6**

Table 4: The experimental results with different percentages of the MixATIS dataset. For each percentage, the upper results are the SLIM baseline, while the bottom ones are our PACL framework. **Bold** numbers indicate the better result for each percentage.
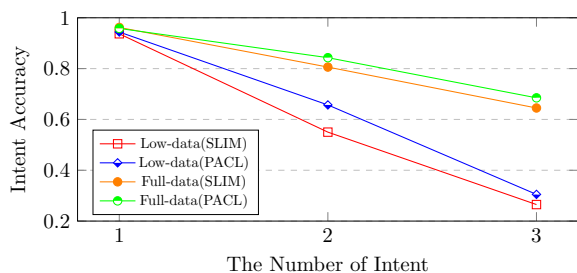

Figure 3: The Intent Accuracy of different numbers of intents trained on low- and high-data of MixATIS dataset.

contrastive learning. Notably, the model performs much worse optimizing 10% of training data than on the other proportions, meanwhile, PACL obtains the maximum performance optimization (8% improvement on Intent Accuracy and 7.2% improvement on Overall Accuracy) on 10% of the data. This also proves that our PACL framework can help the model construct better embedding spaces indirectly, especially for low-data scenarios.

5.3. The Effect of Different Number Intents

Next, we conduct a comparative analysis of the impact of the SLIM baseline and our PACL framework on labels with varying numbers of intents. As shown in Figure 3, it is clear that on single-intent test data, there are minimal disparities in model performance across different methods and data volumes. However, the disparity between the performance of models trained on low- and high-data
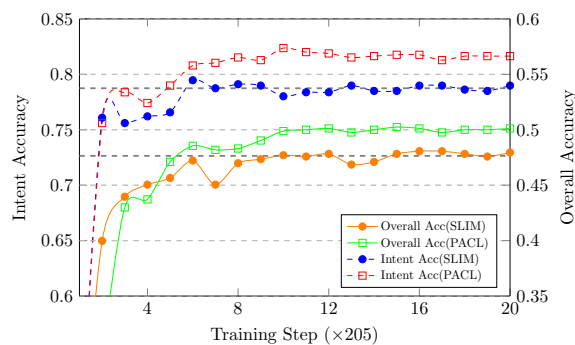

Figure 4: The learning curve of intent accuracy and overall accuracy on MixATIS test set every fixed validation step.

scenarios obtains significant improvements on test data with 2- and 3-intent labels. This divergence can be attributed to the fact that training on samples with multiple intents aids the model in comprehending individual intents, whereas the reverse is not as effective. Moreover, the degree of enhancement achieved by our method for the model intensifies with the increasing number of intents. In low-data scenarios, our approach enhances intent accuracy by 8% for 2-intent test data and by 9% for 3-intent data. In high-data scenarios, our method improves intent accuracy by 3.7% for samples with 2-intent labels and by 4% for samples with 3-intent labels. These results underscore the effectiveness of our PACL framework for multi-intent NLU while enhancing the efficiency of utilizing low-resource data.

5.4. Learning Curves

In order to validate the impact of our PACL framework during the training process, we visually analyze the learning curve of SLIM (Cai et al., 2022a) on the test set of the MixATIS dataset (Figure 4). The intent accuracy of the baseline and PACL approaches is denoted by dashed lines, while solid lines represent their respective overall accuracy curves. Our proposed method demonstrates swifter convergence compared to the baseline, showcasing consistent enhancements in both intent accuracy and overall accuracy. Notably, the overall accuracy of PACL in the early stage is lower than baseline. This makes sense because the pretraining process uses word-level inputs, and the model needs to adapt to the utterance-level pattern in the early stage of training process.

5.5. Visualization

As shown in Figure 5, we utilized the t-SNE algorithm to visualize the distribution of intent embed-
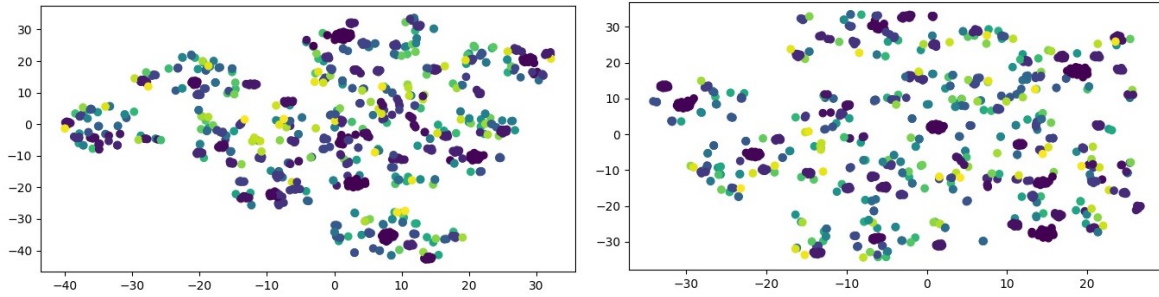
1784

Figure 5: The distribution of intent embeddings in MixATIS dataset. The left one is trained on SLIM baseline, and the other one is trained with our PACL method.
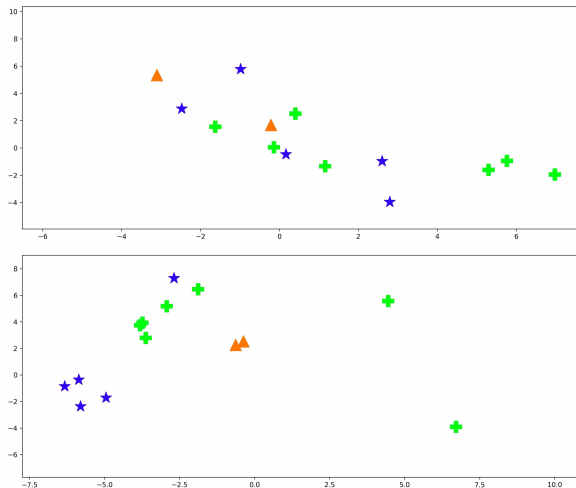

Figure 6: The distribution of several shared intents samples in MixATIS test set. The upper one is original BERT model, while the lower one is the BERT model pre-trained with our word-level pre-training strategy.

5.6. Case Study

In addition to the t-SNE visualization, we exhibit two illustrative examples to validate the performance of our method more specifically, in which example 1 is drawn from the MixATIS dataset, while example 2 is drawn from the MixSNIPS dataset. As depicted in Table 5, the SLIM baseline encounters challenges in making accurate predictions on both multi-intent detection and slot filling tasks, which primarily stems from the high similarity between multiple intents. In contrast, our method successfully demarcates a more distinct boundary, which ameliorates the mispredictions (Example 1) and underpredictions (Example 2) problems. Furthermore, our method addresses inaccuracies in slot predictions. The consistency of slot and intent corrections also demonstrates the strong correlation between the token and its corresponding intent.

6. Conclusion

This paper introduces a novel two-stage prediction-aware contrastive learning framework for multi-intent NLU, which significantly enhances model performance through leveraging word-level pre-training and prediction-aware contrastive fine-tuning. Our method acquires knowledge not only from distinct intents, but also from shared common intents. The experimental results show that our method significantly improved for both low-data and full-data scenarios. As for the future work, we plan to explore how to maximize the impact of contrastive learning with a reduced number of positive samples.

7. Limitations

The main limitation of this approach is training efficiency. Although it will not increase the inference time, it has more training cost because of the gradient calculation for contrastive samples, especially since the number K is strongly related to the model performance. Besides, even though our approach

dings in the MixATIS dataset. The embeddings derived from the SLIM baseline demonstrate the classification boundaries with limited capability to distinguish between shared intent samples. In contrast, the intent embeddings produced by our PACL model exhibit a more distinct distribution, indicating a heightened ability to differentiate between closely related intent classes.

In further detail, we randomly extract several samples with the same shared intents from the MixATIS test dataset to visualize the distinctions in embedding spaces between the original BERT model and the BERT model pre-trained using our word-level pre-training strategy. Notably, the visual analysis in Figure 6 reveals that, following our word-level pre-training strategy, the model exhibits the capacity to distinguish between utterances with distinct intents, even when they share some intents. This enhanced capability positions the model for improved performance during the fine-tuning stage.

	Example 1
Query	list la and also how many Canadian airlines flights use aircraft dh8
Model answer	
Intent	atis_city#atis_quantity
Slots	O B-city_name O O O O B-airline_name I-airline_name O O O B-aircraft_code
Predicted by SLIM (Cai et al., 2022a)	
Intent	atis_abbreviation#atis_quantity
Slots	O B-aircraft_code O O O O B-airline_name I-airline_name O O O B-aircraft_code
Predicted by PACL	
Intent	atis_city#atis_quantity
Slots	O B-city_name O O O O B-airline_name I-airline_name O O O B-aircraft_code
	Example 2
Query	book a restaurant for one person at 7 am and then play the album journeyman
Model answer	
Intent	BookRestaurant#SearchCreativeWork
Slots	O O B-restaurant_type O B-party_size_number O O B-timeRange I-timeRange O O O O B-object_type B-object_name
Predicted by SLIM (Cai et al., 2022a)	
Intent	BookRestaurant#PlayMusic#SearchCreativeWork
Slots	O O B-restaurant_type O B-party_size_number O O B-timeRange I-timeRange O O O O B-music_item B-album
Predicted by PACL	
Intent	BookRestaurant#SearchCreativeWork
Slots	O O B-restaurant_type O B-party_size_number O O B-timeRange I-timeRange O O O O B-object_type B-object_name

Table 5: The examples of how our method solves the intent misclassification problem and improves overall accuracy. Red texts indicates the incorrect results, while blue texts indicates the correct results.

utilizes the relation between the two tasks for associative contrastive learning, only intent labels are used for supervision. It might be better to design a strategy to introduce slot-label supervision.

9. Bibliographical References

Samyadeep Basu, Amr Sharaf, Karine Ip Kiun Chong, Alex Fischer, Vishal Rohra, Michael Amoake, Hazem El-Hammamy, Ehi Nosakhare, Vijay Ramani, and Benjamin Han. 2022. Strategies to improve few-shot learning for intent classification and slot-filling. In *Proceedings of the Workshop on Structured and Unstructured Knowledge Integration (SUKI)*, pages 17–25, Seattle, USA. Association for Computational Linguistics.

Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings. 2022a. Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7607–7611. IEEE.

Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings. 2022b. Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7607–7611. IEEE.

Guanhua Chen, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. 2023. Multi-level curriculum learning for multi-turn dialogue generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3958–3967.

Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022a. Joint multiple intent detection and slot filling via self-distillation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7612–7616.

Lisung Chen, Nuo Chen, Yuexian Zou, Yong Wang, and Xinzhong Sun. 2022b. A transformer-based threshold-free framework for multi-intent NLU.

In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7187–7192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909.

Lizhi Cheng, Weijia Jia, and Wenmian Yang. 2021. An effective non-autoregressive model for spoken language understanding. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 241–250.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5467–5471. Association for Computational Linguistics.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-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 564–569, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Yutai Hou, Yongkui Lai, Cheng Chen, Wanxiang Che, and Ting Liu. 2021. Learning to bridge metric spaces: Few-shot joint learning of intent detection and slot filling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3190–3200, Online. Association for Computational Linguistics.

Byeongchang Kim, Seonghan Ryu, and Gary Geunbae Lee. 2017. Two-stage multi-intent detection for spoken language understanding. *Multimedia Tools and Applications*, 76:11377–11390.

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 2021*, pages 1945–1955, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

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 *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 178–188, Online. Association for Computational Linguistics.

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

Mengxiao Song, Bowen Yu, Li Quangang, Wang Yubin, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot

filling with global intent-slot co-occurrence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7967–7977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nguyen Anh Tu, Hoang Thi Thu Uyen, Tu Minh Phuong, and Ngo Xuan Bach. 2023. Joint multiple intent detection and slot filling with supervised contrastive learning and self-distillation. *CoRR*, abs/2308.14654.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Paweł Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7544–7559, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. SlotRefine: A fast non-autoregressive model for joint intent detection and slot filling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1932–1937, Online. Association for Computational Linguistics.

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 *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7203–7208, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Puyang Xu and Ruhi Sarikaya. 2013. Exploiting shared information for multi-intent natural language sentence classification. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 3785–3789. ISCA.

Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, and Boaz Carmeli. 2023. QAID: question answering inspired few-shot intent detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14310–14318.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5259–5267, Florence, Italy. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

10. Language Resource References

Hou, Yutai and Lai, Yongkui and Wu, Yushan and Che, Wanxiang and Liu, Ting. 2021. *Few-shot learning for multi-label intent detection*. AAAI Press. PID 10.1609/aaai.v35i14.17541.

Qin, Libo and Xu, Xiao and Che, Wanxiang and Liu, Ting. 2020. *AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling*. Association for Computational Linguistics. PID 10.18653/v1/2020.findings-emnlp.163.