HC^{2}L: Hybrid and Cooperative Contrastive Learning for Cross-lingual Spoken Language Understanding

Bowen Xing and Ivor W. Tsang, *Fellow, IEEE*

Abstract—State-of-the-art model for zero-shot cross-lingual spoken language understanding performs *cross-lingual unsupervised contrastive learning* to achieve the label-agnostic semantic alignment between each utterance and its code-switched data. However, it ignores the precious intent/slot labels, whose label information is promising to help capture the label-aware semantics structure and then leverage supervised contrastive learning to improve both source and target languages' semantics. In this paper, we propose Hybrid and Cooperative Contrastive Learning to address this problem. Apart from cross-lingual unsupervised contrastive learning, we design a holistic approach that exploits *source language supervised contrastive learning*, *cross-lingual supervised contrastive learning* and *multilingual supervised contrastive learning* to perform label-aware semantics alignments in a comprehensive manner. Each kind of supervised contrastive learning mechanism includes both single-task and joint-task scenarios. In our model, one contrastive learning mechanism's input is enhanced by others. Thus the total four contrastive learning mechanisms are cooperative to learn more consistent and discriminative representations in the virtuous cycle during the training process. Experiments show that our model obtains consistent improvements over 9 languages, achieving new state-of-the-art performance.

Index Terms—Dialog System, Spoken Language Understanding, Contrastive Learning, Cross-lingual

1 Introduction

Spoken language understanding (SLU) plays a crucial role in task-oriented dialog systems [1], [2], [3], [4]. It includes two subtasks: intent detection, which is a sentence-level classification task, and slot filling, which is a sequence labeling task. Great progress has been achieved in the past decade, while current SLU methods require a large amount of data, which is impractical in some scenarios. To this end, zero-shot cross-lingual spoken language understanding [5], [6], [7], [8] has been explored and attracted increasing interest because it can significantly reduce the effort for data annotation and transfer the task knowledge learned from the high-resource language into the target low-resource language.

Most previous models can only conduct implicit multilingual semantics alignment via parameters sharing [9], [10], [11], [12]. GL-CLEF [8], which is the up-to-date state-of-the-art model, proposes to leverage cross-lingual unsupervised contrastive learning (CL) to perform explicit semantics alignment between the utterance and its multilingual view obtained by code-switching [11]. Its contrastive learning mechanism is conceptually illustrated in Fig. 1 (a). GL-CLEF pulls together the current utterance and its multilingual view in the semantic space. At the same time, the current utterance is pushed apart from all other utterances as well as their multilingual views. Despite the significant improvements that GL-CLEF achieves, we find that it suffers from a drawback: its contrastive learning mechanism does not consider any label information. This leads to two issues. First, some utterances and multilingual views have the same/similar label(s) as the current utterance. Thus, it is intuitive to pull them together. However, they are toughly separated only because they are not the current utterance's multilingual view. Second, some utterances and multilingual views have different labels, so their semantics are supposed to be pushed apart. However, GL-CLEF cannot achieve this due to ignoring the label information.

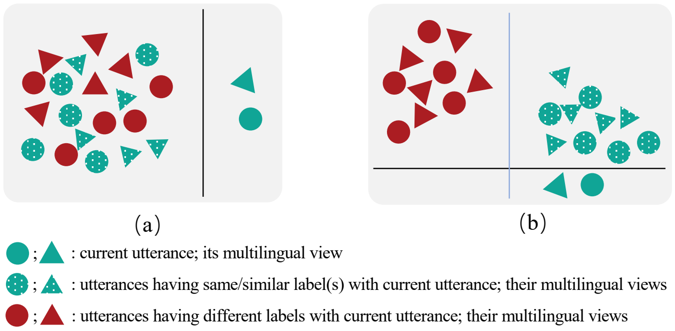

(a)
(b)
● ; ▲ : current utterance; its multilingual view
● ; ▲ : utterances having same/similar label(s) with current utterance; their multilingual views
● ; ▲ : utterances having different labels with current utterance; their multilingual views

Fig. 1. Conceptual comparison of the contrastive learning mechanisms of GL-CLEF (a) and our HC^{2}L (b). The black line denotes the margin caused by the cross-lingual unsupervised contrastive learning proposed by GL-CLEF. The blue line denotes the margin caused by our proposed three kinds of supervised contrastive learning mechanisms.

Therefore, we argue that the precious and sufficient label information is urgent to be leveraged to perform supervised

- *Bowen Xing is with Beijing Key Laboratory of Knowledge Engineering for Materials Science, School of Computer and Communication Engineering, University of Science and Technology Beijing. E-mail: bwxing714@gmail.com*
- *Ivor Tsang is with CFAR, Agency for Science, Technology and Research; IHPC, Agency for Science, Technology and Research; School of Computer Science and Engineering, Nanyang Technological University; Australian Artificial Intelligence Institute, University of Technology Sydney. E-mail: ivor_tsang@cfar.a-star.edu.sg*

TABLE 1
Summary of the four contrastive learning (CL) mechanisms in our model.

CL Mechanisms	Input	Enhancement
Cross-lingual Unsupervised CL	Source Language Semantics; Multilingual View Semantics	Multilingual View Semantics
Source Language Supervised CL	Labels; Source Language Semantics	Source Language Semantics
Cross-lingual Supervised CL	Labels; Source Language Semantics; Multilingual View Semantics	Multilingual View Semantics
Multilingual Supervised CL	Labels; Multilingual View Semantics	Multilingual View Semantics

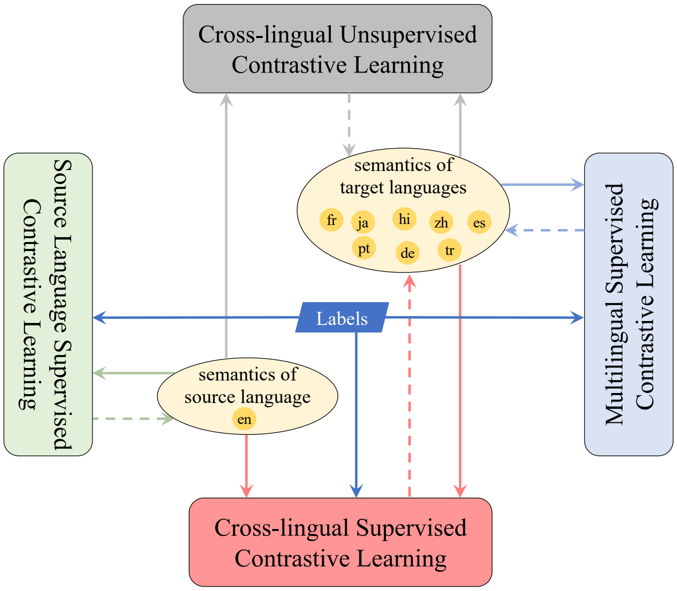

Cross-lingual Unsupervised Contrastive Learning
semantics of target languages
fr
ja
hi
zh
es
pt
de
tr
Source Language Supervised Contrastive Learning
Multilingual Supervised Contrastive Learning
Labels
semantics of source language
en
Cross-lingual Supervised Contrastive Learning

Fig. 2. Illustration of the cooperation of the hybrid contrastive learning mechanisms in our HC^{2}L model. Solid arrows denote input. Dashed arrows denote enhancement.

contrastive learning to capture the label-aware semantics structure. And we discover three promising perspectives for designing supervised contrastive learning: source language, cross-lingual and multilingual. Then we propose HC^{2}L: Hybrid and Cooperative Contrastive Learning to address the above problems, as conceptually illustrated in Fig.1(b). We design three kinds of supervised contrastive learning mechanisms: (1) ***source language supervised contrastive learning***, which enhances the source language semantics via label-aware semantics alignment; (2) ***cross-lingual supervised contrastive learning***, which transfers the knowledge from the source language to target languages via performing the label-aware alignment between the source language semantics and the multilingual view semantics; (3) ***multilingual supervised contrastive learning***, which aligns the multilingual view semantics that has the same/similar label(s). Each of the three kinds of supervised contrastive learning comprehensively includes both of the single-task (intent/slot) and the joint-task (intent+slot) supervised contrastive learning mechanisms. The intent/slot supervised contrastive learning mechanisms leverage intent/slot labels to capture high-level semantic structure via performing label-aware semantics alignment. However, there is no given label for the joint task. To this end, we construct it by ourselves using the given intent and slot labels, and we propose the joint-task multi-label supervised contrastive learning, which can model the dual-task correlations.

To achieve explicit cross-lingual semantics alignment, HC^{2}L also includes the cross-lingual unsupervised contrastive learning proposed by GL-CLEF [8]. As shown in Fig. 2, the total four kinds of contrastive learning mechanisms in our model have mutual influence and interdependencies: the input semantics of one contrastive learning mechanism is reinforced by other contrastive learning mechanisms. In this way, the hybrid contrastive learning mechanisms in our model can cooperate with each other to learn better and better semantic representations in the training procedure.

Thanks to the proposed supervised contrastive learning mechanism, our HC^{2}L model have two advantages over previous models:

- It can learn more *consistent* semantic representations across different languages.
- It can learn more *discriminative* semantics representations across different classes.

We evaluate our model on MultiATIS++, which is a benchmark including 9 different languages. Experiment results show that our model obtains about 10% average improvements over the previous best-performing model on overall accuracy. Further analysis verifies the effectiveness of our proposed contrastive learning mechanisms and it is proven that our proposed supervised contrastive learning mechanisms contribute more than the cross-lingual unsupervised contrastive learning. The visualizations of learned representations show that our model can learned more consistent representations across different languages. And in the same time, our model effectively pushes apart the semantics corresponding to different classes.

2 Related Works

2.1 Zero-shot Cross-lingual Spoken Language Understanding

Spoken language understanding [13], [14], [15], [16] is a core component of dialog systems. It usually includes two subtasks: intent detection and slot filling. Intent detection is a sentence-level classification task aiming to predict the intent expressed in an utterance. Slot filling is a sequence labeling task that assigns a slot label to each word. In recent years, as researchers have realized the correlations between intent detection and slot filling, a group of models are proposed to jointly tackle the two tasks via leveraging the correlative information [3], [4], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. Co-guiding Net [4] makes the first time to model the mutual guidance between the multi-intent

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3402746

detection and slot filling via heterogeneous semantics-label graphs. ReLa-Net [26] exploits the dual-dependencies and leverages them for dual-task interaction and joint decoding.

However, these SLU models largely rely on rich-source training data, which is not always practicable, especially for some low-source languages. To solve this problem, zero-shot cross-lingual SLU is explored and has attracted increasing attention. Since mBERT [29] is a strong baseline for cross-lingual language understanding, some methods are proposed to improve mBERT at the pre-training stage [5], [30], [31], [32], [33], [34].

Besides, some other works aim to perform semantics alignment between the source language and target languages at the fine-tuning stage. Attention-informed mixed-language training [9] proposes code-mixing to construct multilingual training samples containing phrases from both of source and target languages. And CoSDA [11] further proposes multilingual code-switching to better perform multilingual semantics alignment. GL-CLEF [8], which is the up-to-date state-of-the-art, proposes to perform explicit multilingual semantics alignment via cross-lingual unsupervised CL. LAJ-MCL [35] proposes to model the utterance-slot-word structure by a multi-level contrastive learning framework. FC-MTLF [36] proposes to leverage the neural machine translation task to improve cross-lingual SLU. DiffSLU [37] leverages a powerful diffusion model to enhance the mutual guidance between the slot and intent. Differently, we make the first attempt to perform three kinds of label-aware semantics alignments via source language, cross-lingual and multilingual supervised CL.

2.2 Contrastive Learning for NLP

Contrastive learning has been widely adopted to improve representations in NLP tasks [38], [39], [40], [41]. In the natural language inference task, pairwise supervised CL [38] bridges semantic entailment and contradiction understanding with high-level categorical concept encoding. Hierarchy-guided contrastive learning [41] directly incorporates the hierarchy into the text encoder for hierarchical text classification. In this paper, we propose three kinds of supervised CL (e.g., source language supervised CL, cross-lingual supervised CL and multilingual supervised CL) to achieve comprehensive label-aware semantics alignments for zero-shot cross-lingual spoken language understanding .

3 HC^{2}L

Zero-shot cross-lingual SLU aims to train the model in a source language (e.g., English) and then directly apply it to target languages (e.g., French, Japan, Chinese) for testing. And there are two subtasks:

- Intent detection. It is a sentence-level classification problem aiming to predict the correct intent label l^I.
- Slot filling. It is a token-level sequence labeling task mapping the input utterance word sequence $X = \{x_1, ..., x_n\}$ to the slot sequence $\{l_1^s, ..., l_n^s\}$, where n denotes the word number.

In this section, we introduce our HC^{2}L model in detail. Its architecture is shown in Fig. 3, and its four contrastive learning mechanisms are summarized in Table 1.

3.1 Backbone Framework

3.1.1 Encoder

Following state-of-the-art method [8], we adopt the pre-trained mBERT model to encode the input utterance word sequence, and the representation of the first sub-token of a word is used for the word representation. Then we obtain the word hidden states:

$$\mathbf{H} = \{h_{\texttt{CLS}}, h_1, ..., h_n\} \tag{1}$$

where `[cls]` is the special token at the beginning of the input sequence and $h_{[\texttt{CLS}]}$ is taken as the sentence representation; h_t denotes the first sub-token representation of word x_t. Then we follow [8] to generate the multilingual code-switched data (multilingual view), which is fed to Multilingual BERT (mBERT) [29] to generate the hidden states:

$$\mathbf{H}^{\text{ml}} = \{h_{\texttt{CLS}}^{\text{ml}}, h_1^{\text{ml}}, ..., h_n^{\text{ml}}\}. \tag{2}$$

3.1.2 Intent Detection Decoder

. The sentence representation $h_{\text{CLS}}^{\text{ml}}$ is fed to a softmax classifier to predict the intent label:

$$l^I = \text{softmax}(W_I \; h_{\texttt{CLS}}^{\text{ml}} + b_I) \tag{3}$$

where W_I and b_I denote weight matrix and bias.

3.1.3 Slot Filling Decoder

. For word x_t, we feed its representation h_t^{ml} into the slot classifier to predict its slot label:

$$l_t^s = \text{softmax}(W_S \; h_t^{\text{ml}} + b_S) \tag{4}$$

where W_S and b_S denote weight matrix and bias.

3.2 Sample Queues

Since our model performs both unsupervised and supervised contrastive learning, inspired by [42], we maintain a set of sample queues that store not only the previously encoded features but also their labels.
1. Utterance representation queue: $Q^u = \{h_{\texttt{CLS}}^k\}_{k=1}^K$. It stores the sentence representations of previously encoded source language utterances, where K is the length of the queue.
2. Word representation queue: $Q^w = \{h_1^k, ..., h_n^k\}_{k=1}^K$. It stores the word representation sequence of previously encoded source language utterances.
3. Multilingual view utterance representation queue: $Q_{ml}^u = \{h_{\texttt{CLS}}^{[\text{ml},k]}\}_{k=1}^K$. It stores the sentence representations of previously encoded code-switched utterances (multilingual views).
4. Multilingual view word representation queue: $Q_{ml}^w = \{h_1^{[\text{ml},k]}, ..., h_n^{[\text{ml},k]}\}_{k=1}^K$. It stores the word representation sequence of previously encoded code-switched utterances (multilingual views).
5. Intent label queue: $Q_l^I = \{\hat{l}_k^I\}_{k=1}^K$. It stores the one-hot intent labels of previously encoded utterances.
6. Slot label queue: $Q_l^S = \{\hat{l}_{[k,1]}^s, ..., \hat{l}_{[k,n]}^s\}_{k=1}^K$ It stores the one-hot slot label sequence of previously encoded utterances.

The queues are updated with the current batch's features and labels while dequeuing the oldest ones. And each utterance and its multilingual view share the same intent label and slot labels. For instance, $h_{\texttt{CLS}}^k$ and $h_{\texttt{CLS}}^{[\text{ml},k]}$ correspond to the same one-hot intent label $\hat{l}_k^I$; h_1^k and $h_1^{[\text{ml},k]}$ correspond to the same one-hot slot label $\hat{l}_{[k,1]}^s$.

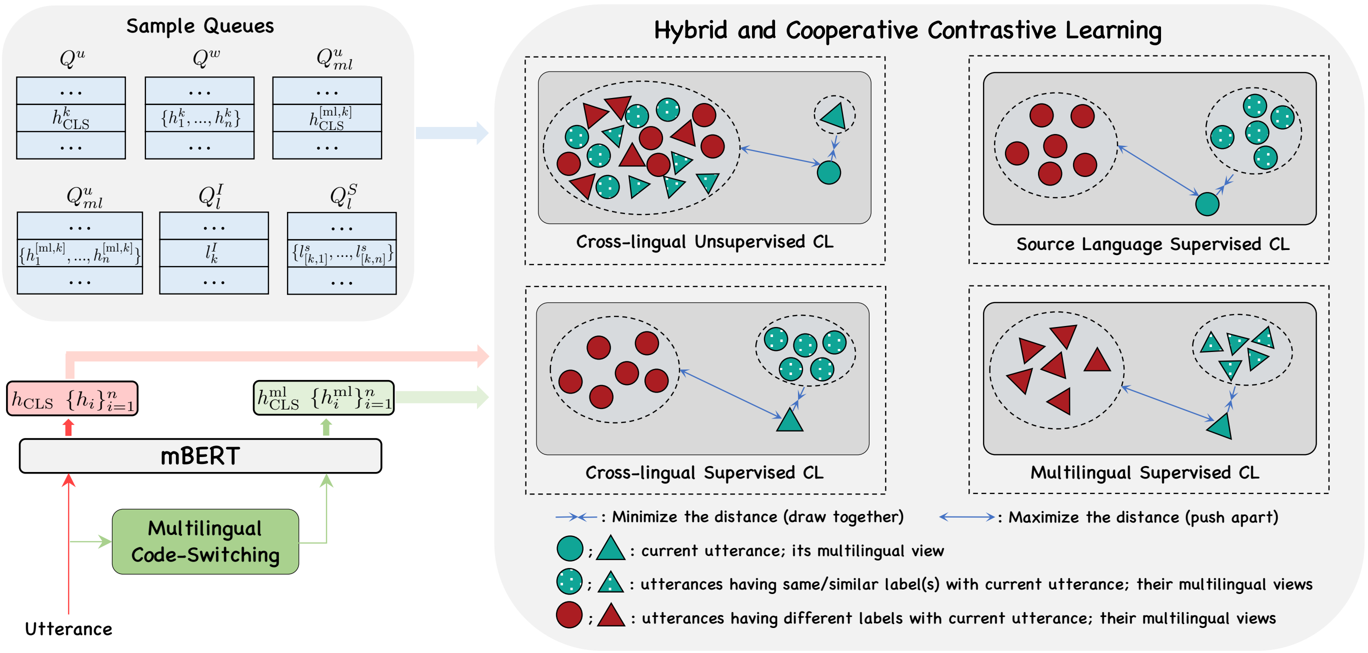

Sample Queues
Q^u
Q^w
Q^u_{ml}
h^k_{CLS}
$\{h^k_1, ..., h^k_n\}$
$h^{[\text{ml},k]}_{\text{CLS}}$
Q^u_{ml}
Q^I_l
Q^S_l
$\{h^{[\text{ml},k]}_1, ..., h^{[\text{ml},k]}_n\}$
l^I_k
$\{l^s_{[k,1]}, ..., l^s_{[k,n]}\}$
h_{CLS} $\{h_i\}^n_{i=1}$
$h^{\text{ml}}_{\text{CLS}}$ $\{h^{\text{ml}}_i\}^n_{i=1}$
mBERT
Multilingual Code-Switching
Utterance
Hybrid and Cooperative Contrastive Learning
Cross-lingual Unsupervised CL
Source Language Supervised CL
Cross-lingual Supervised CL
Multilingual Supervised CL
: Minimize the distance (draw together)
: Maximize the distance (push apart)
; : current utterance; its multilingual view
; : utterances having same/similar label(s) with current utterance; their multilingual views
; : utterances having different labels with current utterance; their multilingual views

Fig. 3. The architecture of our HC^{2}L model. For simplicity, the SLU prediction module is omitted, and our three supervised contrastive learning mechanisms are illustrated in single-label manner.

3.3 Cross-lingual Unsupervised Contrastive Learning

Our model includes cross-lingual unsupervised contrastive learning, whose effectiveness has been verified in [8]. The sentence and word representations of the current utterance's multilingual view are positive samples, while all other representations in Q^u, Q^w, Q^u_{ml} and Q^w_{ml} are negative samples.

Intent. The cross-lingual intent unsupervised contrastive learning mechanism aims to align the sentence representations of the current utterance and its multilingual view. Specifically, it can be formulated as follows:

$$\mathcal{L}^I_{un} = -\log \frac{e^{\frac{h_{\text{CLS}} \cdot h^{\text{ml}}_{\text{CLS}}}{\tau}}}{e^{\frac{h_{\text{CLS}} \cdot h^{\text{ml}}_{\text{CLS}}}{\tau}} + \sum_k^K [e^{\frac{h_{\text{CLS}} \cdot h^k_{\text{CLS}}}{\tau}} + e^{\frac{h_{\text{CLS}} \cdot h^{[\text{ml},k]}_{\text{CLS}}}{\tau}}]} \tag{5}$$

where τ denotes the temperature.

Slot. As slot filling is token-level, the cross-lingual slot unsupervised contrastive learning mechanism aims to perform token-level semantics alignment. Specifically, it can be formulated as follows:

$$\mathcal{L}^S_{un} = -\frac{1}{n^2} \sum_i^n \sum_j^n \log \frac{e^{\frac{h_i \cdot h^{\text{ml}}_j}{\tau}}}{e^{\frac{h_i \cdot h^{\text{ml}}_j}{\tau}} + \sum_k^K [e^{\frac{h_i \cdot h^k_j}{\tau}} + e^{\frac{h_i \cdot h^{[\text{ml},k]}_j}{\tau}}]} \tag{6}$$

Intent-Slot. The cross-lingual global intent-slot unsupervised contrastive learning mechanism aims to model the global semantic alignment for both intent and slot. Specifically, it can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\text{GIS}}_{\text{un}} = &-\frac{1}{n} \sum_j^n \log \frac{e^{\frac{h_{\text{CLS}} \cdot h_j}{\tau}}}{e^{\frac{h_{\text{CLS}}, h_j}{\tau}} + \sum_k^K [e^{\frac{h_{\text{CLS}}, h^k_j}{\tau}} + e^{\frac{h_{\text{CLS}}, h^{[\text{ml},k]}_j}{\tau}}]} + \\ &- \frac{1}{n} \sum_j^n \log \frac{e^{\frac{h_{\text{CLS}} \cdot h^{\text{ml}}_j}{\tau}}}{e^{\frac{h_{\text{CLS}}, h^{\text{ml}}_j}{\tau}} + \sum_k^K [e^{\frac{h_{\text{CLS}}, h^k_j}{\tau}} + e^{\frac{h_{\text{CLS}}, h^{[\text{ml},k]}_j}{\tau}}]}\end{aligned} \tag{7}$$

The motivation behind this design is that in a single sentence, its slots and intent are usually highly related from the semantics perspective. Therefore, GL-CLEF takes the intent representation (h_{CLS}) in a sentence and its own slots' representations (h_j) to naturally constitute a form of positive pairs, while the slots' representations of other sentences can form negative pairs.

3.4 Source Language Supervised contrastive learning

Regarding the current utterance as the anchor, we propose the source language supervised CL to perform source language label-aware semantics alignment. It draws together the anchor's sentence (word) representation and source language samples in Q^u (Q^w) or pushes them apart regarding whether they have the same/similar label(s).

3.4.1 *Intent Supervised contrastive learning*

Taking the current utterance's sentence representation and the samples in Q^u as input, source language intent supervised contrastive learning pulls together the anchor's sentence representation and the ones of positive samples, while distinguishing the anchor's sentence representation from the ones of negative samples. And we adopt the cosine function with temperature τ' to measure the similarity of the two representations:

$$s(a,b) = \frac{a^T \cdot b}{\|a\| \cdot \|b\| \cdot \tau'}. \tag{8}$$

The samples sharing the same intent label with the anchor is regarded as positive samples, while other ones are negative samples. However, the positive samples and negative samples are mixed together in Q^u. To this end, we propose to use the Hadamard product of the current utterance's one-hot intent label and the queue sample's one to automatically retrieve the positive samples. Denoting current utterance as i, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{I}}_{\mathrm{slscl}} &= -\sum_{k}^{K} \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \log \frac{e^{s(h^i_{\mathrm{CLS}}, h^k_{\mathrm{CLS}})}}{\sum_{j}^{K} e^{s(h^i_{\mathrm{CLS}}, h^j_{\mathrm{CLS}})}} \\ \mu_{ik} &= \hat{l}^I_i \odot \hat{l}^I_k\end{aligned} \tag{9}$$

where μ_{ik} equals 0 or 1, indicating whether the k-th sample in Q^u is a positive sample; $\odot$ denotes Hadamard product.

3.4.2 *Slot Supervised Contrastive Learning*

It aims to align the source language's word representations that have the same slot label. Similarly, we adopt the Hadamard product of one-hot slot labels to automatically retrieve positive word representation samples from Q^w. This contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{S}}_{\mathrm{slscl}} &= -\frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} \sum_{k}^{K} \frac{\mu_i^{[k,j]}}{\sum_{a}^{K} \mu_{ia}} \log \frac{e^{s(h_i, h_j^k)}}{\sum_{a}^{K} e^{s(h_i, h_j^a)}} \\ \mu_i^{[k,j]} &= \hat{l}^s_i \odot \hat{l}^s_{[k,j]}\end{aligned} \tag{10}$$

where $\mu_i^{[k,j]}$ equals 0 or 1, indicating whether the j-th word representation of the k-th sample in Q^w is a positive sample of the current utterance's i-th word representation.

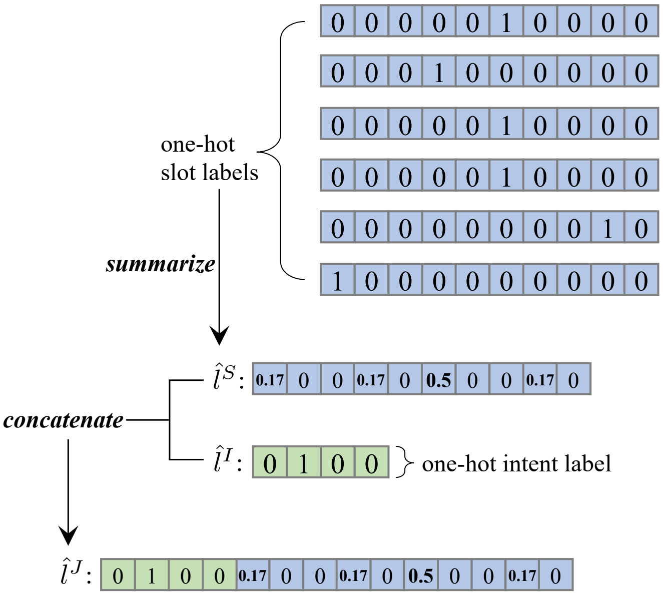
3.4.3 *Joint-task Multi-Label Supervised Contrastive Learning*

To jointly model intent detection and slot filling in supervised contrastive learning, we have to construct the sentence-level joint-task label by ourselves, because it is not provided in the dataset. To this end, we first obtain the sentence-level slot label $\hat{l}^S$ by summarizing all non-O slot labels in one label vector:

$$\hat{l}^S = \frac{\sum_{i=1, l_i^s \neq \mathrm{O}}^{N} \hat{l}_i^s}{\sum_{i=1, l_i^s \neq \mathrm{O}}^{N} 1} \tag{11}$$

where l_i^s denotes the slot label if i-th word, and $\hat{l}_i^s$ denotes the one-hot label. $\hat{l}^S$ can represent the slot-specific semantics of the utterance. Then we concatenate $\hat{l}^S$ and the one-hot intent label $\hat{l}^I$ to form the joint-task label $\hat{l}^J$. An example of obtaining $\hat{l}^J$ is shown in Fig. 4.

Since $\hat{l}^J$ is not a one-hot label, this contrastive learning mechanism performs multi-label supervised contrastive learning. In this case, some samples may share some common labels with the current utterance, while others may share all labels or no one. Therefore, how to measure the golden similarity of the two instances is the key challenge. In this paper, we propose to use the Hadamard product with normalization to achieve this. Denoting the current utterance as i, the whole process of this contrastive learning mechanism can be formalized as follows:



0 0 0 0 0 1 0 0 0 0
0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 0 1 0
1 0 0 0 0 0 0 0 0 0

$\hat{l}^S$: 0.17 0 0 0.17 0 0.5 0 0 0.17 0

$\hat{l}^I$: 0 1 0 0

$\hat{l}^J$: 0 1 0 0 0.17 0 0 0.17 0 0.5 0 0 0.17 0

Fig. 4. Illustration of the process of constructing the sentence-level joint-task label using the one-hot intent labels and the one-hot slot labels (non-O). We first summarize the one-hot slot labels into a single sentence-level soft slot label, which is then concatenated with the one-hot intent label to form the final sentence-level joint task label. In this example, w.l.o.g, there are total 4 different intents and 10 different slots.The reason we do not count the O slot is that it does not convey any semantics information.

$$\begin{aligned}\mathcal{L}^{\mathrm{Joint}}_{\mathrm{slscl}} &= -\sum_{k}^{K} w_{ik} \log \frac{e^{s(h^i_{\mathrm{CLS}}, h^k_{\mathrm{CLS}})}}{\sum_{j}^{K} e^{s(h^i_{\mathrm{CLS}}, h^j_{\mathrm{CLS}})}} \\ w_{ik} &= \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \\ \mu_{ij} &= \hat{l}^J_i \odot \hat{l}^J_j\end{aligned} \tag{12}$$

A large μ_{ik} denotes the sample k is quite similar to the anchor and leads to a large w_{ik} assigned to the loss function to pull them closer. Instead, $\mu_{ik} = 0$ denotes they have totally different labels. Since their distance only appears in the denominator, they will be pushed apart by the negative gradient.

3.5 Cross-lingual Supervised Contrastive Learning

Generally, the model can learn relatively high-quality source language representations from the training set. To transfer the label-aware semantics knowledge from the source language to the target languages, we propose the cross-lingual supervised contrastive learning mechanism to perform explicit label-aware cross-lingual semantics alignment. The current utterance's multilingual view is regarded as the anchor, and its sentence (word) representations are drawn together with the source language positive samples in Q^u (Q^w) that share the same/similar label(s), while they are distinguished from the negative samples that have different labels.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2024.3402746

3.5.1 Intent Supervised Contrastive Learning

Denoting current utterance as i, similar to Eq. 4, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{I}}_{\mathrm{clscl}} &= -\sum_{k}^{K} \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \log \frac{e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{k}_{\mathrm{CLS}})}}{\sum_{k}^{K} e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{k}_{\mathrm{CLS}})}} \\ \mu_{ik} &= \hat{l}^{I}_{i} \odot \hat{l}^{I}_{k}\end{aligned} \tag{13}$$

where μ_{ik} equals 0 or 1, indicating whether the k-th sample in Q^u is a positive sample of the current utterance's sentence representation. $h^{[\mathrm{ml},i]}_{\mathrm{CLS}}$ denotes the sentence representation of the current utterance's multilingual view.

3.5.2 Slot Supervised Contrastive Learning

Similar to Eq. 5, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{S}}_{\mathrm{clscl}} &= -\frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} \sum_{k}^{K} \frac{\mu_{i}^{[k,j]}}{\sum_{a}^{K} \mu_{ia}} \log \frac{e^{s(h^{\mathrm{ml}}_{i}, h^{k}_{j})}}{\sum_{a}^{K} e^{s(h^{\mathrm{ml}}_{i}, h^{a}_{j})}} \\ \mu_{i}^{[k,j]} &= \hat{l}^{s}_{i} \odot \hat{l}^{s}_{[k,j]}\end{aligned} \tag{14}$$

where $\mu_{i}^{[k,j]}$ equals 0 or 1, indicating whether the j-th word representation of the k-th sample in Q^w is a positive sample of the current utterance's i-th word representation. h^{ml}_{i} denotes the i-th word representation of the current utterance's multilingual view.

3.5.3 Joint-task Multi-Label Supervised Contrastive Learning

Similar to Eq. 6, this contrastive learning mechanism can be formulated as:

$$\begin{aligned}\mathcal{L}^{\mathrm{Joint}}_{\mathrm{clscl}} &= -\sum_{k}^{K} w_{ik} \log \frac{e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{k}_{\mathrm{CLS}})}}{\sum_{j}^{K} e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{j}_{\mathrm{CLS}})}} \\ w_{ik} &= \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \\ \mu_{ij} &= \hat{l}^{J}_{i} \odot \hat{l}^{J}_{j}\end{aligned} \tag{15}$$

3.6 Multilingual Supervised Contrastive Learning

Although we cannot directly perform supervised contrastive learning on the target languages due to the lack of training data, we propose the multilingual supervised contrastive learning to achieve the pseudo target language supervised contrastive learning inspired by the two facts: (1) we have the representations of the multilingual view, which contains the semantics of target languages; (2) multilingual views also have ground-truth labels because each one shares the same labels with its original source language utterance. This contrastive learning mechanism aims to pull the current multilingual view's sentence (word) representation together with the samples in Q^u_{ml} (Q^w_{ml}) or push them apart regarding whether they share the same/similar label(s).

3.6.1 Intent Supervised Contrastive Learning

Denoting the current source language utterance as i, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{I}}_{\mathrm{mlscl}} &= -\sum_{k}^{K} \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \log \frac{e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{[\mathrm{ml},k]}_{\mathrm{CLS}})}}{\sum_{j}^{K} e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{[\mathrm{ml},j]}_{\mathrm{CLS}})}} \\ \mu_{ik} &= \hat{l}^{I}_{i} \odot \hat{l}^{I}_{k}\end{aligned} \tag{16}$$

where μ_{ik} equals 0 or 1, indicating whether the k-th sample in Q^u_{ml} is a positive sample of the current utterance's multilingual view's sentence representation.

3.6.2 Slot Supervised Contrastive Learning

Similar to Eq. 5, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{S}}_{\mathrm{mlscl}} &= -\frac{1}{n^2} \sum_{i}^{n} \sum_{j}^{n} \sum_{k}^{K} \frac{\mu_{i}^{[k,j]}}{\sum_{a}^{K} \mu_{ia}} \log \frac{e^{s(h^{\mathrm{ml}}_{i}, h^{[\mathrm{ml},k]}_{j})}}{\sum_{a}^{K} e^{s(h^{\mathrm{ml}}_{i}, h^{[\mathrm{ml},a]}_{j})}} \\ \mu_{i}^{[k,j]} &= \hat{l}^{s}_{i} \odot l^{s\,\hat{}}_{[k,j]}\end{aligned} \tag{17}$$

where $\mu_{i}^{[k,j]}$ equals 0 or 1, indicating whether the j-th word representation of the k-th sample in Q^w_{ml} is a positive sample of the current utterance's multilingual view's i-th word representation.

3.6.3 Joint-task Multi-Label Supervised Contrastive Learning

Denoting the current source language utterance as i, this contrastive learning mechanism can be formulated as follows:

$$\begin{aligned}\mathcal{L}^{\mathrm{Joint}}_{\mathrm{mlscl}} &= -\sum_{k}^{K} w_{ik} \log \frac{e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{[\mathrm{ml},k]}_{\mathrm{CLS}})}}{\sum_{j}^{K} e^{s(h^{[\mathrm{ml},i]}_{\mathrm{CLS}}, h^{[\mathrm{ml},j]}_{\mathrm{CLS}})}} \\ w_{ik} &= \frac{\mu_{ik}}{\sum_{j}^{K} \mu_{ij}} \\ \mu_{ij} &= \hat{l}^{J}_{i} \odot \hat{l}^{J}_{j}\end{aligned} \tag{18}$$

3.7 Training Objective

Denoting $\mathcal{L}_I$ and $\mathcal{L}_S$ as the standard loss function for intent detection and slot filling, the final training objective of HC^{2}L is the weighted sum of $\mathcal{L}_I$, $\mathcal{S}$ and all the above contrastive learning objectives:

$$\begin{aligned}\mathcal{L} &= \lambda_I \mathcal{L}_I + \lambda_S \mathcal{L}_S + \mathcal{L}_{\mathrm{un}} + \mathcal{L}_{\mathrm{slscl}} + \mathcal{L}_{\mathrm{clscl}} + \mathcal{L}_{\mathrm{mlscl}} \\ \mathcal{L}_{\mathrm{un}} &= \lambda^{I}_{\mathrm{un}} \mathcal{L}^{I}_{un} + \lambda^{\mathrm{S}}_{\mathrm{un}} \mathcal{L}^{\mathrm{S}}_{\mathrm{un}} + \lambda^{\mathrm{GIS}}_{\mathrm{un}} \mathcal{L}^{\mathrm{GIS}}_{\mathrm{un}} \\ \mathcal{L}_{\mathrm{slscl}} &= \beta_I \mathcal{L}^{\mathrm{I}}_{\mathrm{slscl}} + \beta_S \mathcal{L}^{\mathrm{S}}_{\mathrm{slscl}} + \beta_J \mathcal{L}^{\mathrm{Joint}}_{\mathrm{slscl}} \\ \mathcal{L}_{\mathrm{clscl}} &= \gamma_1 (\beta_I \mathcal{L}^{\mathrm{I}}_{\mathrm{clscl}} + \beta_S \mathcal{L}^{\mathrm{S}}_{\mathrm{clscl}} + \beta_J \mathcal{L}^{\mathrm{Joint}}_{\mathrm{clscl}}) \\ \mathcal{L}_{\mathrm{mlscl}} &= \gamma_2 (\beta_I \mathcal{L}^{\mathrm{I}}_{\mathrm{mlscl}} + \beta_S \mathcal{L}^{\mathrm{S}}_{\mathrm{mlscl}} + \beta_J \mathcal{L}^{\mathrm{Joint}}_{\mathrm{mlscl}})\end{aligned} \tag{19}$$

where λ_*, β_* and γ_* are hyper-parameters balancing the loss terms. The standard loss functions for intent detection ($\mathcal{L}_I$) and slot filling ($\mathcal{L}_S$) in Eq.13 are defined as following:

$$\begin{aligned}\mathcal{L}_I &= -\sum_{j=1}^{C_I} \hat{\mathbf{y}}^{I}[j] \log \left(L^{I}[j] \right) \\ \mathcal{L}_S &= -\sum_{i=1}^{n} \sum_{j=1}^{C_S} \hat{\mathbf{y}}^{s}_{i}[j] \log (l^{s}_{i}[j])\end{aligned} \tag{20}$$

where C_I and C_S are the sets of intent labels and slot labels, respectively; $\hat{\mathbf{y}}^I$ and $\hat{\mathbf{y}}^s$ are the ground-truth intent labels and slot labels, respectively.

TABLE 2
Main results. Our model significantly outperforms baselines with $p < 0.05$ under the t-test.

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
mBERT [10]	-	95.27	96.35	95.92	90.96	79.42	94.96	69.59	86.27	-
mBERT [29]	98.54	95.40	96.30	94.31	82.41	76.18	94.95	75.10	82.53	88.42
Ensemble-Net [12]	90.26	92.50	96.64	95.18	77.88	77.04	95.30	75.04	84.99	87.20
CoSDA [11]	95.74	94.06	92.29	77.04	82.75	73.25	93.05	80.42	78.95	87.32
GL-CLEF [8]	98.54	98.09	97.91	97.72	86.34	80.02	96.41	81.82	88.24	91.68
HC^{2}L (ours)	**99.10**	**98.88**	**99.02**	**99.12**	**90.37**	**88.83**	**97.87**	**89.23**	**93.06**	**95.05**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
mBERT [10]	-	82.61	74.98	75.71	31.21	35.75	74.05	23.75	62.27	-
mBERT [29]	95.11	80.11	78.22	82.25	26.71	25.40	72.37	41.49	53.22	61.66
Ensemble-Net [12]	85.05	82.75	77.56	76.19	14.14	9.44	74.00	45.63	37.29	55.78
CoSDA [11]	92.29	81.37	76.94	79.36	64.06	66.62	75.05	48.77	77.32	73.47
GL-CLEF [8]	95.89	84.29	85.76	85.85	65.55	66.36	81.50	68.34	78.30	79.09
HC^{2}L (ours)	**96.18**	**90.17**	**87.86**	**88.03**	**71.91**	**75.78**	**83.49**	**71.29**	**81.96**	**82.96**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
mBERT [29]	87.12	52.69	52.02	37.29	4.92	7.11	43.49	4.33	18.58	36.29
AR-S2S-PTR [7]	86.83	34.00	40.72	17.22	7.45	10.04	33.38	–	23.74	-
IT-S2S-PTR [6]	87.23	39.46	50.06	46.78	11.42	12.60	39.30	–	28.72	-
CoSDA [11]	77.04	57.06	46.62	50.06	26.20	28.89	48.77	15.24	46.36	44.03
GL-CLEF [8]	88.69	66.26	63.71	60.05	26.76	32.84	60.54	30.35	53.08	53.56
HC^{2}L (ours)	**89.92**	**72.20**	**66.05**	**67.51**	**34.83**	**42.44**	**63.34**	**36.50**	**58.01**	**58.98**

4 Experiments

4.1 Settings

4.1.1 Dataset

Following previous works, we evaluate our model on the multilingual benchmark dataset of MultiATIS++ [10]. This dataset includes 9 languages: English (en), Spanish (es), Portuguese (pt), German (de), French (fr), Chinese (zh), Japanese (ja), Hindi (hi) and Turkish (tr). The dataset includes 18 kinds of intents and 84 kinds of slot labels. The training set is in English, consisting of 4488 samples. The validation set for each language includes 490 samples, except for Hindi and Turkish, which have 160 and 60 validation samples, respectively. The testing set for each language includes 893 samples, except for Turkish, which has 715 testing samples.

4.1.2 Implementation

Following previous models, we adopt the base case mBERT as the encoder. (e.g., learning rate, batch size, dropout rate, sample queue size, λ_I, λ_S, λ_{un}^I, λ_{un}^S and $\lambda_{\text{un}}^{\text{GIS}}$). We set the learning rate as 5e-6. The batch size is 32. Dropout rate is 0.1. λ_I and λ_S are set as 1. λ_{un}^I, λ_{un}^S and $\lambda_{\text{un}}^{\text{GIS}}$ are set as 0.01, 0.005 and 0.01, respectively. For fair comparisons, all the above hyper-parameters are set as the same as GL-CLEF [8]. In our experiments, we only tune β_I, β_S, β_J, γ_1 and γ_2, which balance the loss terms corresponding to the three kinds of supervised CL proposed in this paper. The hyper-parameters tested in training our models are listed in Table 3. We test all combinations of them and choose the one achieving the highest average of all the 9 languages' overall accuracies.

We select the best-performing model on the validation set and report its performance on the test set. All experiments in this work are conducted on a single NVIDIA A100 80G GPU.

4.1.3 Baseline

We compare our model with the following baselines:
(1) mBERT [29]. It is based on the same model architecture as BERT [29] and adopts the same training procedure. Its training data covers the Wikipedia pages of 104 languages with a shared subword vocabulary. Therefore, mBERT can obtain the share embeddings across languages, which facilitate cross-lingual NLP tasks.
(2) Ensemble-Net [12]. The final predictions of this model are determined by 8 independent models through majority voting. Each model is separately trained on a single source language.
(3) CoSDA [11]. This model proposes a novel data augmentation framework to generate multi-lingual code-switching data that is used to to fine-tune mBERT. CoSDA can align representations from source and multiple target languages.
(4) GL-CLEF [8]. This model proposes to leverage the unsupervised contrastive learning to perform explicit semantics alignment between the utterance and its multilingual view obtained by code-switching.

TABLE 3
Tuned hyper-parameters. Finally, chosen values are in **bold**.

Hyper-parameters	**Values**
β_I	1e-5, 1e-4, 1e-3, **1e-2**, 0.1
β_S	1e-5, **1e-4**, 1e-3, 1e-2, 0.1
β_J	1e-5, **1e-4**, 1e-3, 1e-2, 0.1
γ_1	1e-2, **0.1**, 1.0
γ_2	1e-2, **0.1**, 1.0

For fair comparisons, we reproduce GL-CLEF's results with the default hyper-parameters.

4.2 Main Results

Following previous works, we adopt accuracy, F1 score and overall accuracy for the metrics to evaluate intent

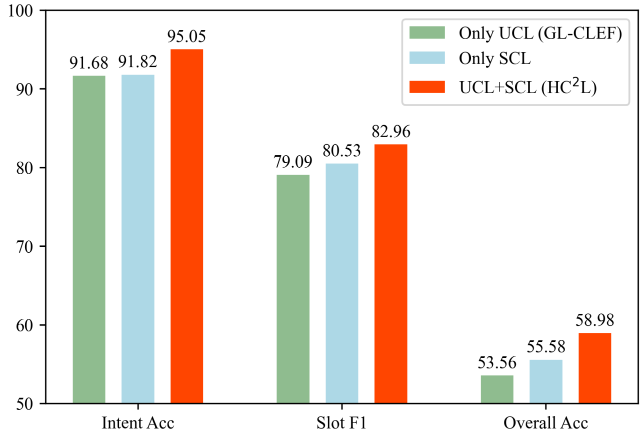


Fig. 5. Ablation results on Only UCL, Only SCL and UCL+SCL. Only UCL is a variant that only performs unsupervised contrastive learning, which is equal to the GL-CLEF model. Only SCL is a variant that only performs the source language, cross-lingual and multilingual supervised contrastive learning. UCL+SCL denotes all unsupervised and supervised contrastive learning mechanism are adopted, which is equal to our HC^{2}L model.

detection, slot filling and sentence-level semantics frame parsing, respectively.

The main results comparison based on mBERT encoder is shown in Table 2. We can observe that our HC^{2}L model significantly outperforms all baselines by a large margin. In terms of overall accuracy, our model obtains a relative improvement of 10.1% over the up-to-date best model GL-CLEF. This demonstrates that our proposed three supervised contrastive learning mechanisms can comprehensively improve the semantics via performing explicit label-aware alignments for the source language, cross-lingual and multilingual scenarios. Besides, we can find that the improvements on languages having inferior performances are sharper. For instance, HC^{2}L achieves nearly 50% relative improvement in terms of overall accuracy for Hindi (hi). This proves that our proposed cross-lingual supervised contrastive learning can effectively transfer the label-aware semantics knowledge from the source language into target languages, and the multilingual supervised contrastive learning can achieve pseudo target language supervised contrastive learning, which can learn better and more discriminative target language semantics.

4.3 Ablation Study

We conduct extensive ablation experiments to study the effect of the contrastive learning mechanisms in our HC^{2}L model.

4.3.1 Effect of Unsupervised Contrastive Learning and Supervised Contrastive Learning

Fig. 5 show the results of the ablation experiments for studying the effect of unsupervised contrastive learning and supervised contrastive learning. Firstly, we can observe that Only SCL outperforms Only UCL on all metrics, proving that our proposed supervised contrastive learning mechanisms contribute more than the cross-lingual unsupervised contrastive learning. This can be attributed to the fact that our proposed three kinds of supervised CL can comprehensively capture the label-aware semantic structure and enhance the label-aware knowledge transfer. Besides, UCL+SCL, namely our HC^{2}L model, achieves the best performance. This proves that the unsupervised contrastive learning and our proposed three kinds of supervised contrastive learning can effectively cooperate with each other to learn better semantic representations in the training procedure.

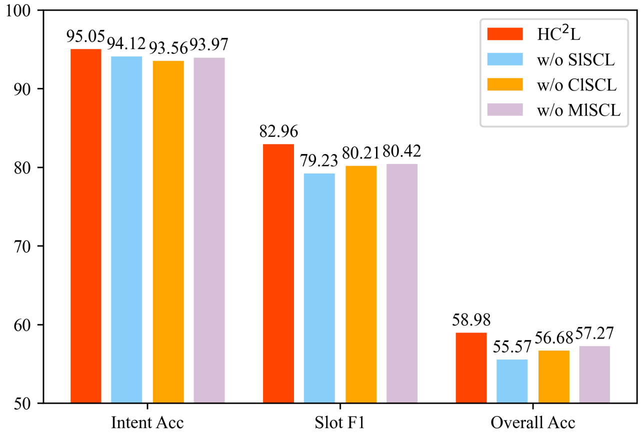


Fig. 6. Ablation results on the three kinds of supervised contrastive learning. SlSCL denotes source language supervised contrastive learning, ClSCL denotes cross-lingual supervised SCL, and MlSCL denotes multilingual supervised contrastive learning.

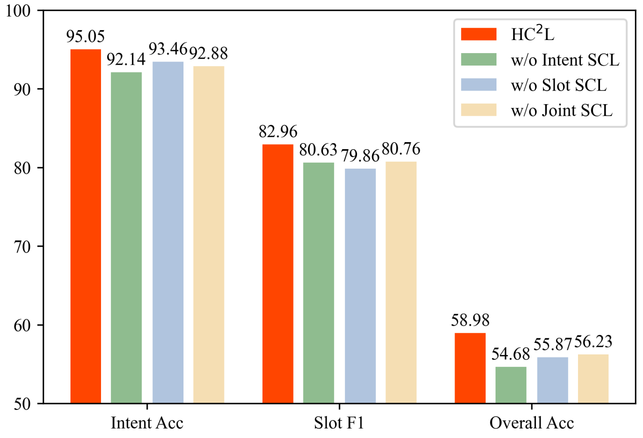
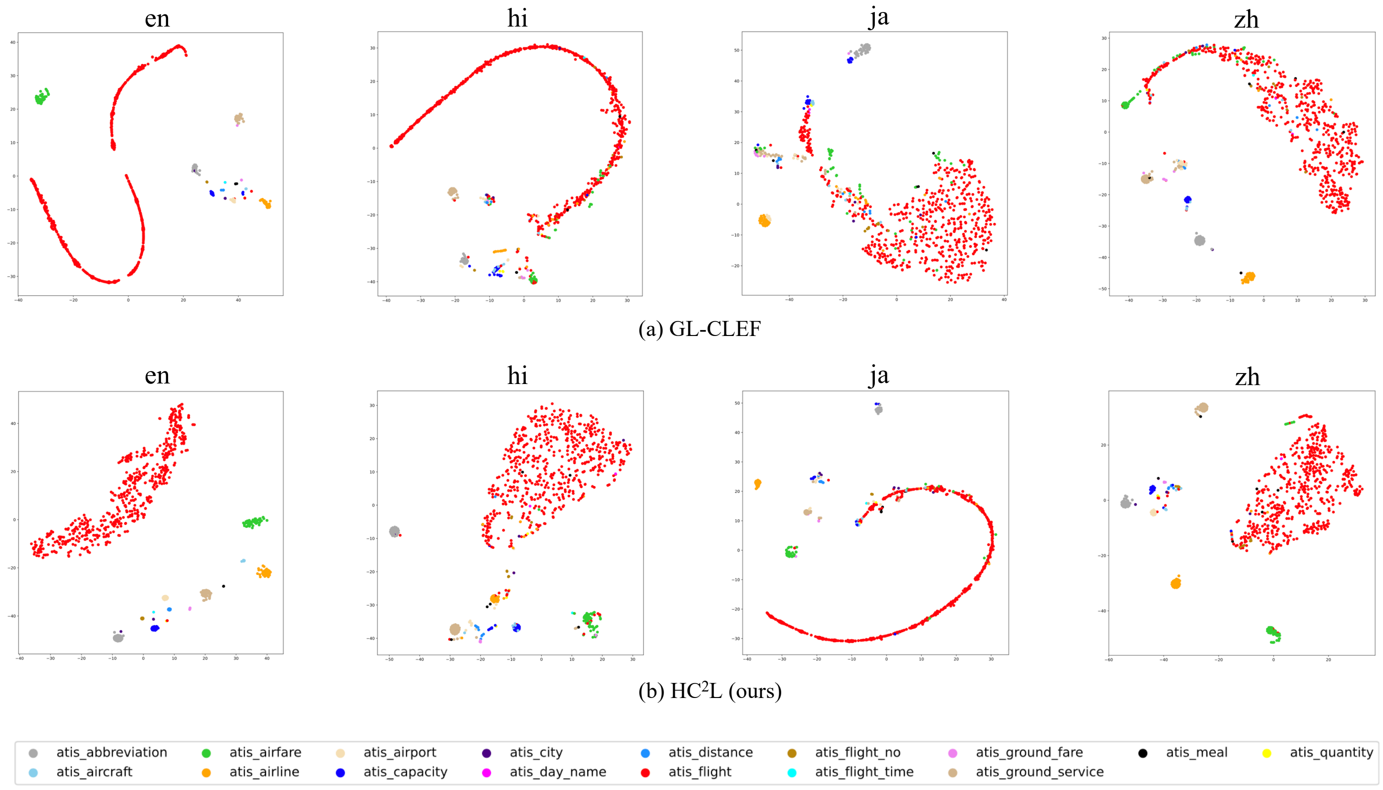


Fig. 7. Ablation results on the single-task supervised contrastive learning and joint-task supervised contrastive learning.

4.3.2 Effect of Different Kinds of Supervised Contrastive Learning

We conduct a group of ablation experiments to study the effectiveness of each kind of supervised CL proposed in this paper. The results are shown in Fig. 6. We can find that all three kinds of supervised contrastive learning bring significant improvements. This is because source language supervised CL can perform explicit label-aware semantics alignments for the source language. And cross-lingual supervised contrastive learning can sufficiently transfer the label-aware semantics knowledge from the source language into target languages. As for multilingual supervised contrastive learning, it can further learn better and more discriminative multilingual representations via drawing together and pushing part the semantics of target languages regarding whether they share the same/similar labels.



(a) GL-CLEF

(b) HC^{2}L (ours)

Fig. 8. TSNE visualization of the sentence embeddings of the source language (en) and three target languages (hi, ja, zh). Different colors denote the sentence embeddings correspond to different intent classes.

TABLE 4
Results based on XLM-R. Our model significantly outperforms baselines with $p < 0.05$ under the t-test.

Intent Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
XLM-R [31]	98.32	97.19	98.03	94.94	88.91	88.50	96.41	72.45	91.15	93.02
XLM-R + GL-CLEF [8]	98.77	97.87	98.28	97.72	85.33	83.52	97.65	85.03	90.59	92.75
XLM-R + HC^{2}L (ours)	**99.22**	**98.43**	**97.42**	**97.72**	**92.61**	**91.20**	**97.98**	**86.15**	**92.95**	**94.85**
Slot F1	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
XLM-R [31]	94.58	72.35	76.72	71.81	60.51	9.31	70.08	45.21	13.44	57.38
XLM-R + GL-CLEF [8]	95.81	87.52	87.51	82.67	67.64	65.15	78.66	55.34	80.69	75.73
XLM-R + HC^{2}L (ours)	**95.83**	**87.32**	**87.76**	**83.30**	**73.65**	**72.84**	**79.24**	**57.12**	**81.32**	**77.74**
Overall Accuracy	en	de	es	fr	hi	ja	pt	tr	zh	Avg.
XLM-R [31]	87.45	43.05	42.93	43.74	19.42	5.76	40.80	9.65	6.60	33.31
XLM-R + GL-CLEF [8]	89.03	68.16	63.96	60.68	30.80	28.10	55.94	18.74	58.79	52.69
XLM-R + HC^{2}L (ours)	**89.36**	**68.16**	**64.70**	**61.57**	**39.20**	**44.13**	**57.51**	**20.98**	**58.90**	**56.06**

4.3.3 Effect of Single-task and Joint-task Supervised Contrastive Learning

We also conduct ablation experiments to study the single-task supervised contrastive learning (Intent SCL and Slot SCL) and the joint-task supervised contrastive learning (Joint SCL). The results are shown in Fig. 7. Intent SCL and Slot SCL use the provided ground-truth labels as the supervision signal to perform supervised contrastive learning. And we can observe that both of them have significant contributions. As for joint-task supervised contrastive learning, since there is no provided label, we construct the sentence-level joint-task labels by ourselves and use them to perform multi-label supervised contrastive learning. The performance gap between the full model (HC^{2}L) and *w/o Joint SCL* demonstrates the effectiveness of Joint SCL, which can capture the dual-task correlations and jointly model the two tasks in supervised contrastive learning for zero-shot cross-lingual SLU.

4.4 Effect of Multilingual Encoder

To verify our method can still work effectively based on other pre-trained multilingual encoders, we conduct experiments based on XLM-R [31] using the same hyper-parameters of mBERT+HC^{2}L. The results are shown in Table 4. We can find that our model can bring a large improvement of 68.3% to XLM-R, and our model significantly outperforms GL-CLEF by 6.4%. This proves that the contributions of our proposed supervised contrastive learning mechanisms are model-agnostic, it can always bring improvements to different multilingual encoders. In the future, if a stronger

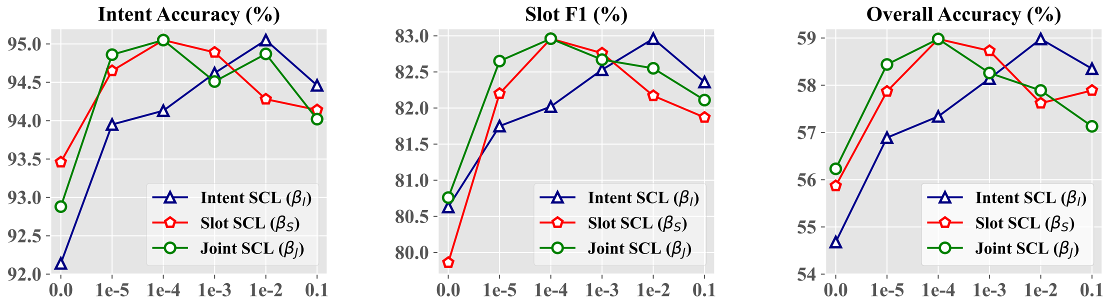


Fig. 9. The performances of our HC^{2}L model on different values of β_I, β_S, β_J.

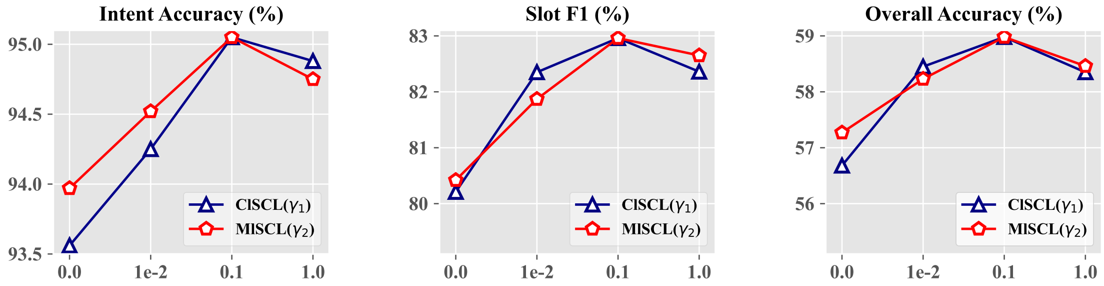


Fig. 10. The performances of our HC^{2}L model on different values of γ_1 and γ_2. CISCL denotes cross-lingual supervised contrastive learning and MISCL denotes multilingual supervised contrastive learning.

pre-trained multilingual encoder is proposed, our model can still be applied to further improve the performance.

4.5 Visualization of Sentence Embeddings

4.5.1 Intent Clustering

We also visualize the sentence embeddings of test samples in different intent classes, as shown in Fig. 8. Compared with GL-CLEF, our HC^{2}L can generate intent clusters exhibiting clearer separation. Although GL-CLEF can disentangle different classes for `en`, our models' intent clusters are clearer and farther away from each other than the ones of GL-CLEF. This can prove that our model can better capture the label-aware semantics structure in the source language, which can be attributed to our proposed source language supervised CL. As for the target languages, our model makes different intent clusters better separated apart than GL-CLEF. Taking `ja` as an example, we can find that compared to the generated sentence embeddings of our model, the ones of GL-CLEF can hardly form intent clusters. Especially, the sentence embeddings of GL-CLEF corresponding to intent 'atis_airfare' and 'atis_flight' obviously overlap with each other, while the sentence embeddings of our model corresponding to intent 'atis_airfare' form a clear cluster and far away from other intent clusters. The above observations prove that our model can generate more discriminative semantics representations for different labels, thanks to our proposed novel supervised CL mechanisms.

4.5.2 Multilingual Clustering

We visualize the sentence embeddings of test samples in all languages, as shown in Fig. 11. GL-CLEF can generally obtain promising representations as different languages' sentence embeddings overlap with each other. However, in Fig. 11, HC^{2}L's red dashed box area is around 12% smaller than GL-CLEF's, indicating that our HC^{2}L generates more tightly entangled sentence embeddings, proving that our model can generate more consistent semantic representations for different languages. This can be attributed to the fact that our model performs explicit label-aware semantics knowledge transfer and label-aware multilingual semantics alignments.

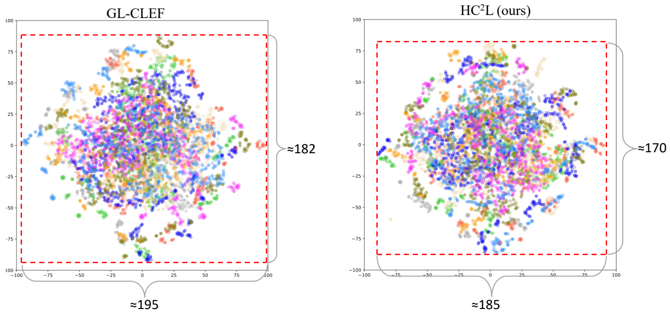


Fig. 11. TSNE visualization of sentence embeddings generated by GL-CLEF and our HC^{2}L model. Different colors denote different languages. All embeddings are normalized on the same scale. A smaller area of the red dashed box denotes the different languages' semantics are more tightly entangled. GL-CLEF's area is 35490 (182x195), while HC^{2}L's area is 31450 (170x185).

4.6 Parameter Analysis

β_I, β_S and β_J control the balance of the single-task supervised CL and joint-task supervised CL. Besides, they determine the value of $\mathcal{L}_{\text{slscl}}$. γ_1 and γ_2 control the extent of cross-lingual supervised CL and multilingual supervised CL. From Fig. 9 and 10, we can find that on each hyper-parameter, the performances first grow and then drop after the peaks. And the best values of β_I, β_S, β_J, γ_1 and γ_2 are 1e-

TABLE 5
Comparison on computation efficiency.

Models	GPU Memory Required	Training Time per Epoch
GL-CLEF	20.7GB	33s
HC^{2}L (ours)	18.9GB	35s

2, 1e-4, 1e-4, 0.1 and 0.1, respectively. Smaller values lead to inferior performances because they make the corresponding supervise signal weak, while larger values also make the performance drop because they harm the balance and dilute other supervise signals.

4.7 Computation Efficiency

We compare our HC^{2}L with the state-of-the-art model GL-CLEF on computation efficiency, as shown in Table 5. Compared with GL-CLEF, our model can decrease the required GPU memory by 8.7%, while costing more 6.1% training time. The GPU memory is dominated by the foundation LLM. The reason why our model occupies slightly less GPU memory is that the queue size of our model is 16, while the queue size of GL- CLEF is 32. For inference, the two models cost the same time because they are based on the same SLU backbone and only perform contrastive learning in the training stage.

5 Conclusion

This paper proposes Hybrid and Cooperative Contrastive Learning (HC^{2}L) for zero-shot cross-lingual SLU. Apart from cross-lingual unsupervised CL, which has been exploited by the state-of-the-art model, we propose source language supervised CL, cross-lingual supervised CL and multilingual supervised CL to perform explicit label-aware semantics alignments. Experiments on 9 languages verify the effectiveness of our model, whose four kinds of CL can cooperate to learn better semantic representations.

Acknowledgments

This work was supported by National Key Research and Development Project of China (No. 2022YFC3502303) and Australian Research Council Grant (No.DP200101328). Ivor W. Tsang was also supported by A*STAR Centre for Frontier AI Research.

References

[1] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[2] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[3] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, "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)*, Aug. 2021, pp. 178–188.

[4] B. Xing and I. Tsang, "Co-guiding net: Achieving mutual guidances between multiple intent detection and slot filling via heterogeneous semantics-label graphs," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Dec. 2022, pp. 159–169.

[5] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Q. Zhu, H. Khan, S. Soltan, S. Rawls, and W. Hamza, "Don't parse, insert: Multilingual semantic parsing with insertion based decoding," in *Proceedings of the 24th Conference on Computational Natural Language Learning*, Nov. 2020, pp. 496–506.

[7] S. Rongali, L. Soldaini, E. Monti, and W. Hamza, "Don't parse, generate! a sequence to sequence archecture for task-oriented semantic parsing," in *Proceedings of The Web Conference 2020*, 2020, pp. 2962–2968.

[8] L. Qin, Q. Chen, T. Xie, Q. Li, J.-G. Lou, W. Che, and M.-Y. Kan, "GL-CLeF: A global–local contrastive learning framework for cross-lingual spoken language understanding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 2677–2686.

[9] Z. Liu, G. I. Winata, Z. Lin, P. Xu, and P. Fung, "Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8433–8440.

[10] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 5052–5063.

[11] L. Qin, M. Ni, Y. Zhang, and W. Che, "Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 3853–3860, main track.

[12] E. Razumovskaia, G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, and I. Vulic, "Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems," *Journal of Artificial Intelligence Research*, vol. 74, pp. 1351–1402, 2022.

[13] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 189–194.

[14] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 2993–2999.

[15] N. T. Vu, P. Gupta, H. Adel, and H. Schütze, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6060–6064.

[16] S.-Y. Su, P.-C. Yuan, and Y.-N. Chen, "How time matters: Learning time-decay attention for contextual spoken language understanding in dialogues," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2133–2142.

[17] B. Liu and I. Lane, "Joint online spoken language understanding and language modeling with recurrent neural networks," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Sep. 2016, pp. 22–30.

[18] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Jun. 2018, pp. 753–757.

[19] C. Li, L. Li, and J. Qi, "A self-attentive model with gate mechanism for spoken language understanding," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Oct.-Nov. 2018, pp. 3824–3833.

[20] H. E, P. Niu, Z. Chen, and M. Song, "A novel bi-directional interrelated model for joint intent detection and slot filling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp. 5467–5471.

[21] Y. Liu, F. Meng, J. Zhang, J. Zhou, Y. Chen, and J. Xu, "CM-net: A novel collaborative memory network 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)*, Nov. 2019, pp. 1051–1060.

[22] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "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)*, Nov. 2019, pp. 2078–2087.

[23] L. Qin, X. Xu, W. Che, and T. Liu, "AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Nov. 2020, pp. 1807–1816.

[24] C. Zhang, Y. Li, N. Du, W. Fan, and P. Yu, "Joint slot filling and intent detection via capsule neural networks," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp. 5259–5267.

[25] D. Wu, L. Ding, F. Lu, and J. Xie, "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)*, Nov. 2020, pp. 1932–1937.

[26] B. Xing and I. Tsang, "Group is better than individual: Exploiting label topologies and label relations for joint multiple intent detection and slot filling," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Dec. 2022, pp. 3964–3975.

[27] B. Xing and I. W. Tsang, "Relational temporal graph reasoning for dual-task dialogue language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 170–13 184, 2023.

[28] ——, "Co-guiding for multi-intent spoken language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2965–2980, 2024.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Jun. 2019, pp. 4171–4186.

[30] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, and M. Zhou, "Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks," 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)*, Nov. 2019, pp. 2485–2494.

[31] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 8440–8451.

[32] J. Yang, S. Ma, D. Zhang, S. Wu, Z. Li, and M. Zhou, "Alternating language modeling for cross-lingual pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9386–9393.

[33] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2021, pp. 483–498.

[34] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, and M. Zhou, "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2021, pp. 3576–3588.

[35] S. Liang, L. Shou, J. Pei, M. Gong, W. Zuo, X. Zuo, and D. Jiang, "Label-aware multi-level contrastive learning for cross-lingual spoken language understanding," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Dec. 2022, pp. 9903–9918.

[36] X. Cheng, W. Xu, Z. Yao, Z. Zhu, Y. Li, H. Li, and Y. Zou, "FC-MTLF: A Fine- and Coarse-grained Multi-Task Learning Framework for Cross-Lingual Spoken Language Understanding," in *Proc. INTERSPEECH 2023*, 2023, pp. 690–694.

[37] T. Mao and C. Zhang, "Diffslu: Knowledge distillation based diffusion model for cross-lingual spoken language understanding," in *Proc. INTERSPEECH 2023*, 2023, pp. 715–719.

[38] D. Zhang, S.-W. Li, W. Xiao, H. Zhu, R. Nallapati, A. O. Arnold, and B. Xiang, "Pairwise supervised contrastive learning of sentence representations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 5786–5798.

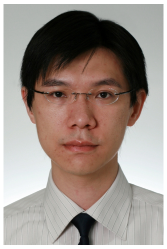
[39] X. Su, R. Wang, and X. Dai, "Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, May 2022, pp. 672–679.

[40] Y. Zhou, P. Liu, and X. Qiu, "KNN-contrastive learning for out-of-domain intent classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 5129–5141.

[41] Z. Wang, P. Wang, L. Huang, X. Sun, and H. Wang, "Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7109–7119.

[42] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

Bowen Xing is an associate professor at school of computer and communication engineering, University of Science and Technology Beijing. He received his B.E. degree and Master degree in computer science from Beijing Institute of Technology in 2017 and 2020, respectively. He obtained his PhD degree in artificial intelligence in 2024, from Australian Artificial Intelligence Institute (AAII), University of Technology Sydney (UTS), under the supervision of Professor Ivor W. Tsang. His research focuses on natural language processing.



Ivor W. Tsang is an IEEE Fellow and the Director of A*STAR Centre for Frontier AI Research (CFAR). Previously, he was a Professor of Artificial Intelligence, at University of Technology Sydney (UTS), and Research Director of the Australian Artificial Intelligence Institute (AAII). His research focuses on transfer learning, deep generative models, learning with weakly supervision, big data analytics for data with extremely high dimensions in features, samples and labels. His work is recognized internationally for its outstanding contributions to those fields. In 2013, Prof Tsang received his ARC Future Fellowship for his outstanding research on big data analytics and large-scale machine learning. In 2019, his JMLR paper "Towards ultrahigh dimensional feature selection for big data" received the International Consortium of Chinese Mathematicians Best Paper Award. In 2020, he was recognized as the AI 2000 AAAI/IJCAI Most Influential Scholar in Australia for his outstanding contributions to the field, between 2009 and 2019. His research on transfer learning was awarded the Best Student Paper Award at CVPR 2010 and the 2014 IEEE TMM Prize Paper Award. In addition, he received the IEEE TNN Outstanding 2004 Paper Award in 2007 for his innovative work on solving the inverse problem of non-linear representations. Recently, Prof Tsang was conferred the IEEE Fellow for his outstanding contributions to large-scale machine learning and transfer learning. Prof Tsang serves as the Editorial Board for the JMLR, MLJ, JAIR, IEEE TPAMI, IEEE TAI, IEEE TBD, and IEEE TETCI. He serves as a Senior Area Chair/Area Chair for NeurIPS, ICML, AAAI and IJCAI, and the steering committee of ACML.