SaLa: Scenario-aware Label Graph Interaction for Multi-intent Spoken Language Understanding

Zhihong Zhu
School of ECE
Peking University
Shenzhen, China
zhihongzhu@stu.pku.edu.cn

Xuxin Cheng
School of ECE
Peking University
Shenzhen, China
chengxx@stu.pku.edu.cn

Zhanpeng Chen
School of ECE
Peking University
Shenzhen, China
troychen927@stu.pku.edu.cn

Zhichang Wang
School of ECE
Peking University
Shenzhen, China
wzcc@stu.pku.edu.cn

Zhiqi Huang
School of ECE
Peking University
Shenzhen, China
zhiqihuang@pku.edu.cn

Yuexian Zou*
School of ECE
Peking University
Shenzhen, China
zouyx@pku.edu.cn

Abstract

Recent joint models for multi-intent detection and slot filling (a.k.a multi-intent SLU) have obtained promising results by leveraging the semantic similarities or co-occurrence relationships between intent and slot labels. However, a critical aspect frequently neglected by current models is the significant correlations between label co-occurrences and specific scenarios, such as watching a movie or booking a ticket, which is essential for understanding user utterances in multi-intent SLU. In this paper, we propose a new framework dubbed SaLa (short for **S**cenario-**a**ware **L**abel gr**a**ph interaction), which effectively captures the dynamic co-occurrence relationships among labels across various scenarios, employing a strategy akin to a divide-and-conquer approach. Concretely, SaLa first autonomously classifies the scenario of utterances, and tracks the co-occurring labels by maintaining a unique co-occurrence matrix for each scenario during the training phase. These scenario-independent co-occurrence matrices are further employed to guide the interactions among label representations through graph propagation to conduct accurate prediction. Extensive experiments on two multi-intent SLU benchmark datasets demonstrate the superiority of our SaLa. More strikingly, SaLa also attains competitive results on four extra single-intent and multi-domain SLU benchmark datasets, demonstrating its strong generalizability.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**.

*Corresponding author.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679676

Keywords

Task-oriented Dialog Systems; Spoken Language Understanding

ACM Reference Format:
Zhihong Zhu, Xuxin Cheng, Zhanpeng Chen, Zhichang Wang, Zhiqi Huang, and Yuexian Zou. 2024. SaLa: Scenario-aware Label Graph Interaction for Multi-intent Spoken Language Understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679676

1 Introduction

Spoken language understanding (SLU) is pivotal for accurately interpreting the user's intent through the construction of semantic frames [49, 53, 54]. In general, SLU encompasses two subtasks: intent detection and slot filling [37]. As illustrated in Figure 1(a), intent detection is a classification task that aims to categorize the intent of user utterances, while slot filling is a sequence labeling task designed to extract relevant semantic concepts.

However, it is common for users to express utterances that encompass multiple intents in real-world scenarios as shown in Figure 1(b), which poses a challenge for single-intent SLU. Recognizing this challenge, multi-intent SLU has been explored [10] and attracted increasing attention. Gangadharaiah and Narayanaswamy [13] makes the first attempt to jointly address multiple intent detection and slot filling within a multi-task learning framework. Due to the high correlations between intents and slots [32], it has become mainstream to study multi-intent SLU jointly, leveraging the inherent dependencies between these two subtasks.

To this end, a bunch of joint models [7, 32, 36, 47, 48] have been proposed to fully mine the correlation among intents and slots. Therein, Qin et al. [33] proposed a model termed AGIF for fine-grained multi-intent detection via graph attention networks (GAT) [39], which adaptively integrates predicted intents into the autoregressive decoding process of slot filling. Based on this, Qin et al. [32] introduced GL-GIN, which builds a local slot-aware graph and a global intent-slot graph for each utterance, obtaining speedup and better performance. Xing and Tsang [46] further proposed Co-guiding Net, which implements a two-stage graph-based framework achieving mutual guidance between intents and slots.

Utterance:	I	need	a	reservation	for	⋯	serves	a	Maple	Bacon	donut	
Slot:	O	O	O	O	O	⋯	O	O	B-sd	I-sd	I-sd	
Intent:	BookRestaurant											(a)
Utterance:	add	David	Axelrod	to	my	futuors	hits	⋯	latin	on	zvooq	
Slot:	O	B-a	l-a	O	O	B-p	l-p	⋯	B-g	O	B-s	
Intent:	AddtoPlaylist, PlayMusic											(b)

Figure 1: Two examples of single-intent (a) and multi-intent (b) spoken language understanding (SLU) in distinct scenarios, where `B/I-sd` denote `B/I-served_dish`, `B/I-a` denote `B/I-artist`, `B/I-p` denote `B/I-playlist`, `B-g` denotes `B-genre`, and `B-s` denotes `B-service`.

To conclude, the aforementioned studies can be reduced to implicit methods, which mostly resort to graph-based models to explore semantic similarities among labels. Additionally, an explicit approach has been introduced, which constructs a co-occurrence probability matrix directly from the whole training data, capturing label co-occurrences at both the corpus level [36] and the utterance level [47]. However, the corpus-level approach aggregates statistics of label co-occurrence across the entire training corpus, which constitutes a coarse approximation and may not precisely capture the subtle interactions between intent and slot labels within individual utterances. Meanwhile, the utterance-level approach encounters difficulties in precisely learning a co-occurrence matrix for each utterance, thereby limiting its effectiveness in accurately facilitating the interactions between intent and slot labels. Consequently, a research question arises: *How can we better model the co-occurrence relationships among labels to achieve accurate prediction?*

Toward this goal, we propose a shift in the granularity of modeling label co-occurrence from the commonly adopted utterance-level or corpus-level to the group-level. This shift is motivated by a fundamental observation: label co-occurrence significantly depends on the scenario, a factor often overlooked in previous multi-intent SLU studies. As illustrated in Figure 1, we expect `AddtoPlaylist` to co-occur with `PlayMusic` in a music management scenario, whereas `BookRestaurant` is more relevant in an online booking scenario. Therefore, we propose to divide the training samples into independent groups according to their scenarios and calculate the label co-occurrence matrix for each group separately. Then, samples within the same group utilize a shared co-occurrence matrix for subsequent feature interactions. Evidently, the obtained group-level label co-occurrence can offer more precise guidance for feature interactions between labels compared to the corpus-level and utterance-level approaches. In this context, two technical challenges remain: *first, how to obtain robust label representations; second, how to determine the scenario to which a given utterance belongs?*

In this paper, as shown in Figure 2, we explore a new **S**cenario-**a**ware **L**abel gr**a**ph interaction framework dubbed SaLa for multi-intent SLU. Specifically, SaLa follows these six steps: (1) For the input utterance, SaLa utilizes a self-attentive encoder to obtain intent-specific and slot-specific representations. (2) SaLa uses a similar encoder structure to obtain the intent and slot label embeddings from the predefined labels. (3) The intent and slot label embeddings are fed into the semantic-attentive label embedder to

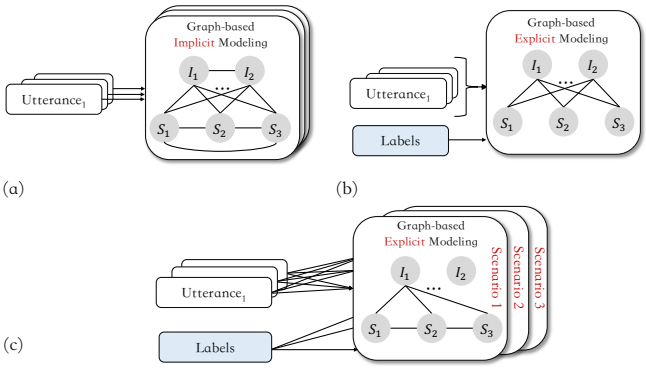


Figure 2: Graph structure comparison among previous works and our approach. Prior research typically constructs a graph for each utterance to model implicit label interactions solely based on utterance semantics (a), or employs a roughly constructed global statistical label graph derived from label co-occurrence in training data applicable to all utterances (b). In contrast, our SaLa systematically models the interaction of label representations for each utterance, guided by the proposed scenario-aware co-occurrence matrices.

fuse the label embeddings with semantic information derived from the utterance. (4) SaLa takes the global representation of the utterance to detect its scenario and updates the corresponding label co-occurrence matrix accordingly. Note that each matrix element represents the number of occurrences of the label pair for the corresponding row and column. They are initially set to zero at the beginning of training and then continuously count the co-occurring labels of utterances by detecting their scenario throughout the training phase. (5) A label graph for each utterance is constructed with labels as nodes and the co-occurrence relationships as edges. The intent and slot label representations are fed into the graph to explore their interactions under the guidance of the scenario-aware label co-occurrence. (6) Finally, two separate decoders are trained for intent detection and slot filling to make predictions.

In a nutshell, the main contributions of this work are three-fold:

- To our best knowledge, this is the first work to explore the correlation between label co-occurrence and scenarios. We propose an effective approach to dynamically model the label co-occurrence for adapting the various dialog scenarios.
- We explore an effective way to integrate utterance representations into label embeddings, and transition the granularity of graph interactions from utterance-level and corpus-level to group-level interactions, achieving a balance between the two.
- Experiment results on two benchmark datasets show that the proposed SaLa framework significantly outperforms previous SOTA models, and further analysis on four additional benchmark datasets verifies the advantages of our SaLa.

2 Problem Formulation

Given the utterance $\mathbf{x}$ consisting of n word tokens $(x_1, x_2, \ldots, x_n)$, the *multiple intent detection* could be formulated as a multi-label classification task which predicts multiple intents $\mathbf{O}^I = (O_1^I, O_2^I, \ldots, O_m^I)$,

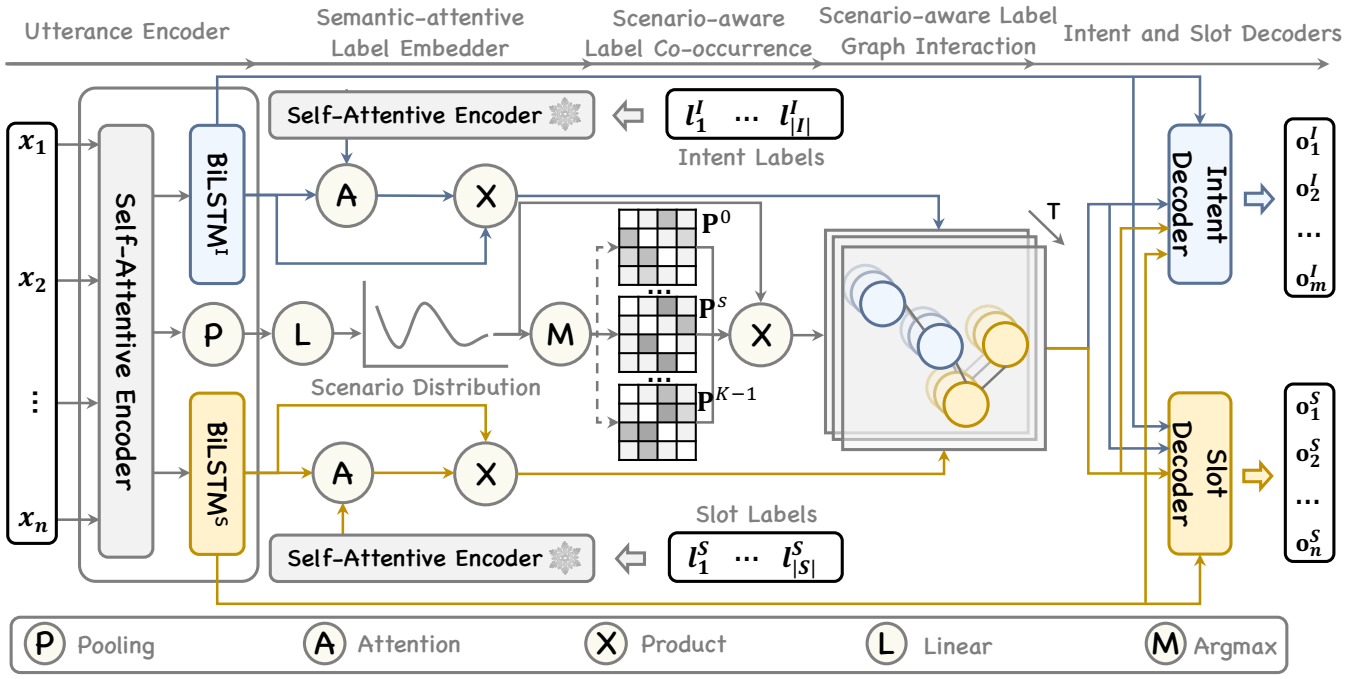


Figure 3: The main architecture of the proposed SALA framework. Better view in color.

where m denotes the number of intents in the input utterance. Meanwhile, *slot filling* can be viewed as a sequence labeling task that predicts a slot label sequence $\mathbf{O}^S = (O_1^S, O_2^S, \ldots, O_n^S)$.

3 Method

In this section, we detail the proposed SALA, whose main architecture is shown in Figure 3. Initially, the SALA takes an utterance and predefined sets of intent and slot labels as inputs, mapping them into utterance and label representations using an utterance encoder (§3.1) and a semantic-attentive label embedder (§3.2), respectively. Next, the scenario-aware label co-occurrence module (§3.3) processes the task-shared utterance representation to detect the scenario of the input utterance, updating the corresponding label co-occurrence matrix with the co-occurring labels. Subsequently, in the scenario-aware label graph interaction module (§3.4), a label graph is constructed to explore the interactions between intent and slot label representations, guided by the scenario-aware label co-occurrence. Finally, two separate decoders (§3.5) are trained for intent and slot prediction. Both intent detection and slot filling are optimized simultaneously through a joint learning scheme (§3.6).

3.1 Utterance Encoder

Following previous works [32, 33, 46], we employ a task-shared encoder and a task-specific encoder as utterance encoder for a fair comparison. Additionally, the discussion using pre-trained language models (PLMs) as the utterance encoder is presented in §4.4.

3.1.1 Task-Shared Encoder. Given an input utterance $\mathbf{x}$ consisting of n word tokens $(x_1, x_2, \ldots, x_n)$, the task-shared encoder creates a vector $\mathbf{e}_i$ to represent the i-th word token x_i by concatenating contextual word embeddings $\mathbf{e}_i^{\text{BiLSTM}}$ and $\mathbf{e}_i^{\text{SA}}$ as follows[1]:

$$\mathbf{e}_i = \mathbf{e}_i^{\text{BiLSTM}} \oplus \mathbf{e}_i^{\text{SA}}, \tag{1}$$

where $\oplus$ denotes concatenation. Here we feed a sequence $\mathbf{e}_{x_1:x_n}$ of real-valued word embeddings into a bidirectional LSTM (BiLSTM) layer [18] and a self-attention (SA) layer [38] to produce the contextual feature vectors $\mathbf{e}_i^{\text{BiLSTM}}$ and $\mathbf{e}_i^{\text{SA}}$, respectively.

3.1.2 Task-Specific Encoder. The task-specific encoder passes the sequence of vectors $\mathbf{e}_{1:n}$ as input to two different single BiLSTM layers to produce task-specific latent vectors $\mathbf{e}_i^I = \text{BiLSTM}^I(\mathbf{e}_{1:n}, i) \in \mathbb{R}^d$ and $\mathbf{e}_i^S = \text{BiLSTM}^S(\mathbf{e}_{1:n}, i) \in \mathbb{R}^d$. These task-specific vectors are concatenated to formulate task-specific matrices $\mathbf{E}^I$ and $\mathbf{E}^S$:

$$\mathbf{E}^I = \left[\mathbf{e}_1^I, \mathbf{e}_2^I, \ldots, \mathbf{e}_n^I\right] \in \mathbb{R}^{n \times d}, \tag{2}$$

$$\mathbf{E}^S = \left[\mathbf{e}_1^S, \mathbf{e}_2^S, \ldots, \mathbf{e}_n^S\right] \in \mathbb{R}^{n \times d}. \tag{3}$$

3.2 Semantic-attentive Label Embedder

Given the pre-defined intent label set $L = \{l_0^I, l_1^I, \ldots, l_{|I|-1}^I\}$ and slot label set $\{l_0^S, l_1^S, \ldots, l_{|S|-1}^S\}$, where $|I|$ and $|S|$ represent the number of intent and slot labels, respectively. We utilize the same encoder structure as described in §3.1.1 to obtain label embeddings. These embeddings are denoted as $\mathbf{H}^\delta \in \mathbb{R}^{|\delta| \times d}$, where $\delta \in \{I, S\}$ (I denotes intent label and S denotes slot label) and $|\delta|$ is the number of intent or slot labels. Notably, these obtained label embeddings are fixed during the training process to prevent overfitting.

[1]For conciseness, the bias terms in this paper are omitted.

Next, we integrate the label embeddings with the semantic information derived from the input utterance. To achieve this, we employ low-rank bilinear pooling [21] to construct an alignment matrix between the task-specific utterance representation and its corresponding label representations. Concretely, we first map the task-specific utterance token $\mathbf{e}_j^\delta$ and corresponding label embedding $\mathbf{h}_i^\delta$ into a task-specific joint embedding space:

$$\mathbf{x}_{ij}^\delta = \mathbf{P}^{\delta\top}(\tanh((\mathbf{U}^{\delta\top}\mathbf{e}_j^\delta) \odot (\mathbf{V}^{\delta\top}\mathbf{h}_i^\delta))), \tag{4}$$

where $\tanh(\cdot)$ denotes the hyperbolic tangent function, $\mathbf{U}^\delta \in \mathbb{R}^{d\times d_\delta}$, $\mathbf{V}^\delta \in \mathbb{R}^{d\times d_\delta}$ and $\mathbf{P}^\delta \in \mathbb{R}^{d_\delta\times d}$ are all learnable parameters. $\odot$ denotes the element-wise multiplication, and d_δ denotes the dimensions of the task-specific joint embedding space.

Then, the normalized attention score α_{ij}^δ for task-specific utterance token $\mathbf{e}_j^\delta$ and label embedding $\mathbf{h}_i^\delta$ is calculated as follows:

$$\alpha_{ij}^\delta = \frac{\exp(\Phi_a^\delta(\mathbf{x}_{ij}^\delta))}{\sum_{j'=0}^{n-1}\exp(\Phi_a^\delta(\mathbf{x}_{ij'}^\delta))}, \tag{5}$$

where $\Phi_a^\delta(\cdot)$ are two learnable feed-forward networks that map the input vectors to logits. Consequently, the semantic-attentive label representation $\mathbf{h}_i^\delta$ is obtained by computing the weighted sum of all token features of the input utterance $\mathbf{x}$ as follows:

$$\mathbf{h}_i^\delta = \sum_{j=0}^{n-1}\alpha_{ij}^\delta \mathbf{e}_j^\delta. \tag{6}$$

3.3 Scenario-aware Label Co-occurrence

3.3.1 Utterance Scenario Detecting. Given that there are K scenarios, which are derived through a statistical analysis of the entire training dataset. To begin, we perform global average pooling along the spatial dimension on the task-shared utterance representation $\mathbf{E}$ to obtain the global representation $\bar{\mathbf{E}}$ of the current input sentence. For PLMs as utterance encoder such as BERT, we directly use the `[CLS]` token for this purpose. Then, the probability that the input utterance $\mathbf{x}$ belongs to the k-th scenario is calculated as follows:

$$\zeta_k = \frac{\exp(\mathbf{w}_k^\top \bar{\mathbf{E}})}{\sum_{k'=0}^{K-1}\exp(\mathbf{w}_k'^\top \bar{\mathbf{E}})}, \tag{7}$$

where $\mathbf{w}_k$ for $k \in \{0, 1, \ldots, K-1\}$ represents a learnable vector, with $\mathbf{w}_k \in \mathbb{R}^d$. This vector serves as a prototype for the k-th scenario, facilitating the clustering of related utterances. Utilizing the derived probability distribution of scenarios, we identify the scenario of the utterance as the one corresponding to the highest probability:

$$s = \underset{k\in\{0,1,\ldots,K-1\}}{\arg\max}\ \zeta_k. \tag{8}$$

In this fashion, the s-th scenario is assigned to the input utterance $\mathbf{x}$, and the corresponding intent and slot labels are utilized to update the co-occurrence matrix for the s-th scenario accordingly.

3.3.2 Label Co-occurrence Modeling. As mentioned above, we aim to mine the group-level label co-occurrence relationships. To this end, a label co-occurrence frequency matrix is maintained for each scenario, which tracks co-occurring labels of the input utterance according to its detected scenario during the training process.

To be specific, if the labels of utterance $\mathbf{x}$ are paired with each other, they are considered as the co-occurrence labels of the s-th scenario. For convenience, all the intents and slot labels are denoted as a multi-hot vector $\mathbf{y} = \left[y_0^I, y_1^I, \ldots, y_{|I|-1}^I, y_0^S, y_1^S, \ldots, y_{|S|-1}^S\right]^\top$, where $y_i \in \{0, 1\}$ for $i \in \{0, 1, \ldots, |I| + |S| - 1\}$ is a binary indicator. And we also temporarily omit the superscript δ to distinguish the intent and the slot in the following part, since they are denoted by the same vector $\mathbf{y}$. Therein, $y_i = 1$ if the label ℓ_i presents in the utterance $\mathbf{x}$ and 0 otherwise. Notably, we do not count slot '`O`' due to its inclusion might introduce noise to the graphs. Thus, the label co-occurrence frequency matrix for the s-th scenario is updated as:

$$\mathbf{C}^s = \mathbf{C}^s + \mathbf{y}\mathbf{y}^\top, \tag{9}$$

where $\mathbf{C}^s \in \mathbb{R}^{(|I|+|S|)\times(|I|+|S|)}$ represent the globally maintained frequency matrix for the s-th scenario throughout the entire training process, which is initialized to zero. In $\mathbf{C}^s$, the diagonal element c_{ii}^s records the frequency of the label ℓ_i, while the off-diagonal element c_{ij}^s captures the frequency of co-occurrences between the label ℓ_i and ℓ_j. Therefore, the probability that label ℓ_j appears in the utterance in the presence of the label ℓ_i in the s-th scenario is:

$$P_{ij}^s = \frac{c_{ij}^s}{c_{ii}^s}. \tag{10}$$

As the training progresses, each co-occurrence frequency matrix continuously updates to reflect the co-occurring labels within the respective scenario, leading the co-occurrence probability matrix to eventually converge towards a steady distribution.

3.4 Scenario-aware Label Graph Interaction

In this subsection, we update the semantic-attentive label representations in a graph propagation manner under the guidance of the scenario-aware label co-occurrence probability matrix. To be specific, we calculate the global co-occurrence probability matrix of the input utterance $\mathbf{x}$ with the obtained co-occurrence probability matrices $\{\mathbf{P}^0, \mathbf{P}^1, \ldots, \mathbf{P}^{K-1}\}$ in §3.3.2 and the predicted scenario probability distribution ζ_k in §3.3.1 of the input utterance $\mathbf{x}$:

$$\mathbf{P}^{\mathbf{x}} = \sum_{k=0}^{K-1}\zeta_k \mathbf{P}^k. \tag{11}$$

Subsequently, we formulate a directed label graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the node set $\mathcal{V}$ represents intent and slot labels, and the edge set $\mathcal{E}$ captures the co-occurrence relationships between adjacent nodes. The edge weights are intuitively initialized based on the label co-occurrence probability matrix $\mathbf{P}^{\mathbf{x}}$. Following this, message passing occurs across the graph $\mathcal{G}$, facilitating the learning of contextual representations for all nodes. Specifically, the feature vector of node v_i at the t-th iteration is represented as $\mathbf{h}_i^t$, which is initially set using the semantic-attentive representation of label ℓ_i, i.e., $\mathbf{h}_i^0 = \mathbf{h}_i$ (*cf.* Eq.(6)). The message that node v_i receives from its neighboring nodes at the t-th iteration is computed as follows:

$$\mathbf{h}_i^t = \sum_{j=0, j\neq i}^{|I|+|S|-1} P_{ji}^{\mathbf{x}}\mathbf{h}_j^{t-1}. \tag{12}$$

This propagation process is iterated T times, allowing the label representations to fully interact with each other. In this fashion,

the vector $\mathbf{h}_i^T$ for node v_i encapsulates both the intrinsic features of label ℓ_i and the contextual information from other labels. This process yields the updated intent and slot label representations $\mathbf{H}^{[T,\delta]}$, which are then used for the final prediction.

3.5 Intent and Slot Decoders

Inspired by Qin et al. [30], we extend a feed-forward network to implicitly integrate intent and slot information. Specifically, we first apply an attention mechanism to extract relevant intent and slot information from the task-specific utterance features $\mathbf{E}^{\delta}$:

$$\mathbf{E}^{\delta} = \text{Softmax}(\mathbf{E}^{\delta}\mathbf{H}^{[T,\delta]^{\top}})\mathbf{H}^{[T,\delta]} + \mathbf{E}^{\delta}. \tag{13}$$

Subsequently, the intent and slot information are integrated through concatenation, denoted as $\mathbf{E_o} = \mathbf{E}^I \oplus \mathbf{E}^S$. To enhance this representation, we incorporate word neighbor features [51] for each token, capturing the context from both preceding and following tokens, which is formulated as follows:

$$\hat{\mathbf{e}}_{\mathbf{o}}^t = \hat{\mathbf{e}}_{\mathbf{o}}^{t-1} \oplus \hat{\mathbf{e}}_{\mathbf{o}}^t \oplus \hat{\mathbf{e}}_{\mathbf{o}}^{t+1}. \tag{14}$$

Next, we combine the label and context features $\hat{\mathbf{E}}_{\mathbf{o}} = \left[\hat{\mathbf{e}}_{\mathbf{o}}^1, \ldots, \hat{\mathbf{e}}_{\mathbf{o}}^n\right]$, and add it to $\mathbf{E}^{\delta}$ to derive the enhanced label information:

$$\begin{aligned} \mathbf{S} &= \text{ReLU}(\hat{\mathbf{E}}_{\mathbf{o}}\mathbf{W}_{\mathbf{o}}^{[1,S]})\mathbf{W}_{\mathbf{o}}^{[2,S]} + \mathbf{E}^S, \\ \mathbf{I} &= \text{ReLU}(\hat{\mathbf{E}}_{\mathbf{o}}\mathbf{W}_{\mathbf{o}}^{[1,I]})\mathbf{W}_{\mathbf{o}}^{[2,I]} + \mathbf{E}^I, \end{aligned} \tag{15}$$

in which $\mathbf{W}_{\mathbf{o}}^{[*,\delta]}$ are trainable parameters.

Now, $\mathbf{I}$ can be used for multiple intent detection:

$$\mathbf{I} = \text{Sigmoid}(\text{ReLU}(\mathbf{I}\mathbf{W}_{\mathbf{o}}^{[3,I]})\mathbf{W}_{\mathbf{o}}^{[4,I]}). \tag{16}$$

The predicted sentence-level intents $\mathbf{O}^I$ are obtained by voting mechanism [32], which can be formulated as follows:

$$\mathbf{O}^I = \{\mathbf{O}_k^I | (\sum_{t=1}^{n} \mathbb{1}[\mathbf{I}_{[t,k]} > 0.5]) > \frac{n}{2}\}, \tag{17}$$

where $\mathbf{I}_{[t,k]}$ represents the prediction probability of token t for the intent $\mathbf{o}_k^I$.

Similar to Eq. 16, $\mathbf{S}$ is used for slot prediction:

$$\mathbf{S} = \text{Softmax}(\text{ReLU}(\mathbf{S}\mathbf{W}_{\mathbf{o}}^{[3,S]})\mathbf{W}_{\mathbf{o}}^{[4,S]}). \tag{18}$$

Finally, the output $\mathbf{O}^S = \text{argmax}(\mathbf{S})$ are the predicted slots sequence of the input utterance X.

3.6 Training Objective

Following previous works, the training objective $\mathcal{L}_S$ of slot filling and the training objective $\mathcal{L}_I$ of intent detection are:

$$\mathcal{L}_S \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{N_S} \hat{\mathbf{O}}_j^{[i,S]} \log\left(\mathbf{O}_j^{[i,S]}\right), \tag{19}$$

$$\mathcal{L}_I \triangleq -\sum_{j=1}^{n}\sum_{i=1}^{N_I} \text{CE}(\hat{\mathbf{O}}_j^{[i,I]}, \mathbf{O}_j^{[i,I]}), \tag{20}$$

$$\text{CE}(\hat{\mathbf{O}}, \mathbf{O}) = \hat{\mathbf{O}}\log(\mathbf{O}) + (1-\hat{\mathbf{O}})\log(1-\mathbf{O}), \tag{21}$$

Dataset	MixATIS	MixSNIPS
Intent categories	18	7
Slot categories	117	72
Training set size	13,162	39,776
Validation set size	756	2,198
Test set size	828	2,199

Table 1: Dataset statistics.

where $\hat{\mathbf{O}}_j^{[i,S]}$ is the gold slot label, $\hat{\mathbf{O}}_j^{[i,I]}$ is the gold intent label, N_S is the number of slot labels, and N_I is the number of intent labels. The final training objective $\mathcal{L}$ is as follows:

$$\mathcal{L} = \lambda\mathcal{L}_I + (1-\lambda)\mathcal{L}_S, \tag{22}$$

where λ represents a hyper-parameter.

4 Experiments

4.1 Datasets and Metrics

Following previous works, we conduct experiments on two benchmark datasets: MixATIS and MixSNIPS [11, 17, 33]. There are 13,162, 756, 828 utterances for training, validation and testing in MixATIS, respectively. MixSNIPS includes 39,776, 2,198, 2,199 utterances for training, validation and testing, respectively. For a fair comparison with previous works, we evaluate the performance of slot filling using F1 score, multiple intent detection using accuracy (Acc), and the sentence-level semantic frame parsing using overall accuracy (Acc) representing all metrics are right in an utterance.

4.2 Implementation Details

Following [32], the word embeddings are trained from scratch, where the dimensions of d is set to 256. For the hyper-parameter λ in Eq.(22), it is set to 0.8 for both MixATIS and MixSNIPS. The graph message propagation times T is empirically set to 3. We adopt AdamW [27] to train SaLa with a learning rate of 1e-3 and a weight decay of 1e-6. The model performing best on the validation set is selected then we report its results on the test set. All experiments are conducted on one single Nvidia V100. The experimental results of our models are averaged over 5 runs with different random seeds.

4.3 Model Zoo

w/o Pre-trained Language Models. We have selected a diverse set of representative and competitive baselines: (i) **Slot-Gated**. [14] introduced a slot-gated joint model designed to learn the correlations between intents and slots. (ii) **Bi-Model**. [41] explored bidirectional interactions between intent detection and slot filling. (iii) **SF-ID Network**. [12] implemented an iterative mechanism to establish a direct connection between intents and slots. (iv) **Stack-Propagation**. [29] utilized a joint model with stack-propagation, where intent detection is used to guide slot filling. (v) **Joint Multiple ID-SF**. [13] proposed a slot-gated mechanism for the joint task of multiple intent detection and slot filling. (vi) **AGIF**. [33] developed an adaptive graph interaction framework to capture fine-grained multi-intent information for slot filling. (vii) **GL-GIN**. [32] introduced a global-local graph interaction network to perform non-autoregressive decoding. (viii) **SDJN**. [7] reformulated multi-intent detection as a weakly supervised task. (ix) **GISCo**. [36] constructed a global

Model	Backbone	MixATIS			MixSNIPS		
		Slot (F1)	Intent (Acc)	Overall (Acc)	Slot (F1)	Intent (Acc)	Overall (Acc)
w/o Pre-trained Language Models							
Slot-Gated$^{\heartsuit}$ [14]	BiLSTM	87.7	63.9	35.5	87.9	94.6	55.4
Bi-Model$^{\heartsuit}$ [41]	BiLSTM	83.9	70.3	34.4	90.7	95.6	63.4
SF-ID Network$^{\heartsuit}$ [12]	BiLSTM	87.4	66.2	34.9	90.6	95.0	59.9
Stack-Propagation$^{\heartsuit}$ [29]	Self-attentive	87.8	72.1	40.1	94.2	96.0	72.9
Joint Mutiple ID-SF$^{\heartsuit}$ [13]	BiLSTM	84.6	73.4	36.1	90.6	95.1	62.9
AGIF$^{\heartsuit}$ [33]	Self-attentive	86.7	74.4	40.8	94.2	95.1	74.2
GL-GIN$^{\heartsuit}$ [32]	Self-attentive	88.3	76.3	43.5	94.9	95.6	75.4
SDJN$^{\heartsuit}$ [7]	Self-attentive	88.2	77.1	44.6	94.4	96.5	75.7
GISCo$^{\heartsuit}$ [36]	Self-attentive	88.5	75.0	48.2	95.0	95.5	75.9
Co-guiding Net$^{\heartsuit}$ [46]	Self-attentive	89.8	79.1	51.3	95.1	97.7	77.5
ReLa-Net$^{\heartsuit}$ [47]	Self-attentive	90.1	78.5	52.2	94.7	97.6	76.1
DARER [48]	Self-attentive	89.2	77.3	49.0	94.9	96.7	76.1
SaLa (Ours)	Self-attentive	**91.6**‡	**82.1**‡	**54.7**‡	**96.5**‡	**98.9**‡	**79.3**‡
w/ Pre-trained Language Models							
LR-Transformer$^{\heartsuit}$ [8]	Transformer	88.0	76.1	43.3	94.4	95.6	74.9
SSRAN$^{\heartsuit}$ [9]	Transformer	89.4	77.9	48.9	95.8	98.4	77.5
SLIM$^{\heartsuit}$ [1]	BERT	88.5	78.3	47.6	96.5	97.2	84.0
DGIF$^{\heartsuit}$ [55]	BERT	88.5	83.3	50.7	95.9	97.8	84.3
TFMN$^{\heartsuit}$ [6]	BERT	88.0	79.8	50.2	96.4	97.7	84.7
UGEN$^{\heartsuit}$ [45]	T5	89.2	83.0	55.3	95.0	96.9	78.8
RoBERTa + AGIF$^{\diamond}$	RoBERTa	86.3	80.1	48.4	95.2	96.8	81.7
RoBERTa + GL-GIN$^{\diamond}$	RoBERTa	86.8	80.6	49.8	95.9	97.2	82.2
RoBERTa + GISCo$^{\diamond}$	RoBERTa	87.3	81.0	52.5	97.0	97.2	82.6
RoBERTa + Co-guiding Net$^{\diamond}$	RoBERTa	88.7	83.2	54.4	96.9	98.0	83.9
RoBERTa + ReLa-Net$^{\diamond}$	RoBERTa	89.2	82.3	54.9	96.7	97.8	83.5
ChatGPT$^{\diamond}$ [28]	-	43.9	65.1	12.8	58.2	94.0	28.9
RoBERTa + SaLa (Ours)	RoBERTa	**90.7**‡	**84.8**‡	**57.0**‡	**97.4**‡	**98.5**†	**85.0**‡

Table 2: Main results (%). $^{\heartsuit}$: results from the corresponding paper. $^{\diamond}$: results by our implementation. Bold: best result, underlined: second best result. † (resp. ‡): SaLa significantly outperforms baselines with $p < 0.05$ (resp. 0.01) under paired t-test.

graph based on inter-label statistical dependencies. (x) **Co-guiding Net**. [46] proposed a two-stage framework that facilitates mutual guidance between intents and slots. (xi) **ReLa-Net**. [47] leveraged label typologies and relations, representing the most recent state-of-the-art model. (xii) **DARER**. [48] investigated relational temporal graph reasoning for fine-grained temporal modeling among labels.

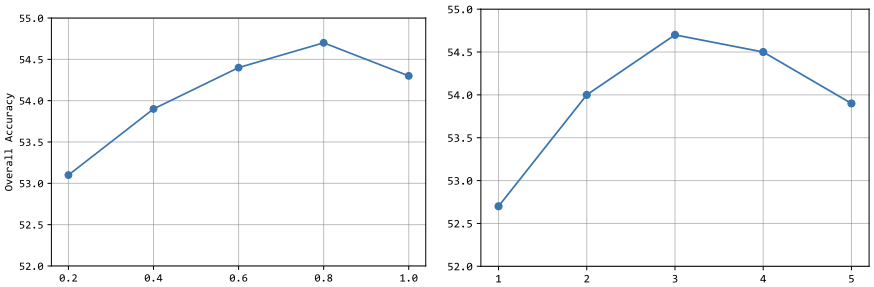
w/ Pre-trained Language Models. To further investigate the potential of SaLa when used alongside pre-trained language models (PLMs), we conducted a comprehensive comparison against existing state-of-the-art baselines that incorporate PLMs: (i) **LR-Transformer**. [8] introduced a layered-refine **Transformer**, featuring a slot label generation task and a layered refinement mechanism. (ii) **SSRAN**. [9] developed a scope-sensitive result attention network based on the **Transformer** architecture to leverage bidirectional interactions between results. (iii) **SLIM**. [1] presented a multi-intent SLU framework utilizing BERT to effectively harness existing annotation data. (iv) **DGIF**. [55] constructed an interactive graph that injects semantic information from labels into node representations. (v) **TFMN**. [6] leveraged the number of intents to achieve threshold-free multi-intent SLU using a **Transformer**-based approach. (vi) **UGEN**. [45] framed the joint multi-intent SLU task as a question-answering problem within a prompt-based paradigm. (vii)–(xi) **RoBERTa + AGIF/GL-GIN/GISCo/Co-guiding Net/ReLa-Net**. We also conducted experiments using five competitive baselines with **RoBERTa$_{\text{base}}$** [25]. (xii) **ChatGPT** (gpt-3.5-turbo-0125). A significant milestone in NLP [28], which we employ in a few-shot learning paradigm as a reference for multi-intent SLU.

4.4 Main Results

The performance of the proposed SaLa and the baselines, both with and without pre-trained language models (PLMs), is presented in Table 2. From this, we can draw the following observations:

Variant	MixATIS			MixSNIPS		
	Slot (F1)	Intent (Acc)	Overall (Acc)	Slot (F1)	Intent (Acc)	Overall (Acc)
SaLa	**91.6** (↓ -)	**82.1** (↓ -)	**54.7** (↓ -)	**96.5** (↓ -)	**98.9** (↓ -)	**79.3** (↓ -)
w/o SaLE	89.5 (↓2.1)	80.4 (↓1.7)	51.9 (↓2.8)	95.6 (↓0.9)	98.3 (↓0.6)	78.6 (↓0.7)
w/o Bilinear Pooling	90.4 (↓1.2)	81.2 (↓0.9)	53.4 (↓1.3)	95.9 (↓0.6)	98.6 (↓0.3)	78.8 (↓0.5)
w/o SaLG	88.8 (↓2.8)	79.5 (↓2.6)	49.0 (↓5.7)	95.3 (↓1.2)	98.0 (↓0.9)	77.9 (↓1.4)
w/ More Parameters	89.5 (↓2.1)	80.2 (↓1.9)	51.3 (↓3.4)	95.5 (↓1.0)	98.2 (↓0.7)	78.4 (↓0.9)

Table 3: Ablation Studies. SaLE: Semantic-attentive Label Embedder. SaLG: Scenario-aware Label Graph Interaction.



(a) Effect of hyper-parameter λ (b) Effect of hyper-parameter T

Figure 4: Analysis experiments of SaLa on the MixATIS dataset (in color). Effect of (a) the trade-off hyper-parameter λ; (b) the graph propagation time T.

(i) ***For models without PLMs on both datasets, SaLa outperforms previous approaches across all metrics by a significant margin***, including implicit graph modeling methods (e.g., **Co-guiding Net**), explicit graph modeling methods using corpus-level label co-occurrence (e.g., **GISCo**), and explicit graph modeling methods using utterance-level label co-occurrence (e.g., **ReLa-Net**). This suggests that SaLa effectively captures correlations between intents and slots through the proposed scenario-aware explicit label graph interaction, thereby enhancing SLU performance. (ii) ***SaLa achieves more substantial improvements on MixATIS***. We hypothesize that this is due to the greater similarity of scenarios in MixATIS. SaLa can optimize distinct label co-occurrence matrices across different scenarios, enabling precise interactions among intent and slot labels. (iii) ***The contributions of SaLa and PLMs are complementary***. From the results with PLMs, we observe that while **RoBERTa** enhances model performance, SaLa combined with **RoBERTa** significantly outperforms other models. This demonstrates the complementary nature of their contributions, indicating that the core strength of SaLa lies in capturing and leveraging label co-occurrence under scenarios to facilitate label interactions, which does not overlap with the advantages of PLMs. (iv) ***Existing joint multi-intent SLU models with PLMs exhibit strong competitiveness***. Even when equipped with a powerful backbone like **T5** [34], the proposed SaLa framework combined with **RoBERTa** delivers superior performance on both benchmarks. (v) ***ChatGPT struggles with multi-intent SLU tasks.*** We applied a method similar to that in [16] to evaluate ChatGPT's performance on these two datasets, using 20 randomly selected examples from the training set. While ChatGPT demonstrates strong few-shot learning abilities in intent detection, it significantly lags behind SaLa in overall accuracy. We suspect this is due to the task-specific knowledge required for this task, which is better captured through fine-tuning. Additionally, the schema of intent and slot labels is complex. Although advanced in-context learning strategies like chain-of-thought may improve ChatGPT's performance to some extent, this is not the focus of our paper. Given ChatGPT's limited success in multi-intent SLU, we argue that designing a robust multi-intent SLU framework remains a challenging and essential task for the NLP community, warranting further exploration.

4.5 Model Analysis

We conduct a set of ablation experiments to verify the advantages of our work, and the results are shown in Table 3.

4.5.1 Effect of Semantic-attentive Label Embedder. To verify the effectiveness of semantic-attentive label embedder, we design a variant termed *w/o* SaLE and its result is shown in Table 3. We can find that employing the same encoder structure of utterance directly for label encoding resulted in a notable decline in overall performance, with decreases of 2.8% and 0.7% on MixATIS and MixSNIPS, respectively. This apparent performance gap between *w/o* SaLE and SaLa underscores the significance of SaLa in capturing the semantics of the input utterance and providing effective representations of initial node features in label co-occurrence graphs. Besides, we substitute the low-rank bilinear pooling with standard attention (Line *w/o* Bilinear Pooling) and observe performance drops across all metrics on both datasets. We attribute this to the effectiveness of

Model	Training Time per Epoch	Latency/Inference Time per Utterance
Co-guiding Net	70s	2.9ms
ReLa-Net	74s	3.0ms
GISCo	76s	3.0ms
SaLa (Ours)	83s	3.1ms

Table 4: Comparison on training and inference time.

Model	ATIS	SNIPS
Stack-Propagation [29]	86.5	86.9
Graph-LSTM [50]	87.6	89.7
Co-Interactive [30]	87.4	90.3
HAN [2]	88.7	91.8
SaLa (Ours)	**89.3**†	**92.5**‡

Table 5: Overall accuracy (%) on ATIS [17] and SNIPS [11] (single-intent SLU) benchmark datasets.

Model	MTOD	ASMixed
Shared-LST [15]	88.7	76.7
Separated-LSTM [15]	89.7	79.5
Multi-Domain adv [24]	88.8	79.5
One-Net [22]	89.4	78.3
Locale-agnostic-Universal [23]	88.5	79.4
Coach [26]	89.5	81.5
Qin et al. [31]	91.3	84.8
SaLa (Ours)	**93.6**‡	**87.0**‡

Table 6: Overall accuracy (%) on MTOD [35] and ASMixed [31] (multi-domain SLU) benchmark datasets.

low-rank bilinear pooling in aligning the utterance with the label space, which enbales the fusion of utterance semantics into label features, thereby enhancing subsequent interactions.

4.5.2 Effect of Scenario-aware Label Graph Interaction. One of the core contributions of our work is achieving precise label graph interaction across different scenarios, while previous works only model a single implicit or explicit graph. To verify its effectiveness, we design a variant termed *w/o* SaLG and its result is shown in Table 3. We observe that overall accuracy drops by 5.7% on MixATIS and 1.4% on MixSNIPS. This proves that the scenario-based label co-occurrence relationships can effectively model precise label interactions between intents and slots, boosting multi-intent SLU systems. Moreover, we replace multiple LSTM layers (2-layers) following [32] as the SaLG to verify that the proposed SaLa rather than the added parameters works. Table 3 (Line *w/* More Parameters) shows the results. We observe that our model outperforms more parameters by 3.4% and 0.9% overall accuracy in two datasets, which shows that the improvements come from the proposed SaLa rather than the involved parameters.

4.6 Hyper-parameters Sensitivity

We conducted a hyperparameter analysis to assess the sensitivity of several key parameters within SaLa: (i) The parameter λ indicates the importance of $\mathcal{L}_I$ between the two tasks. We evaluate the scale range setting $\lambda \in [0.2, 1.0]$ as shown in Figure 4(a). We find that overall accuracy is improved and saturated with $\lambda = 0.8$. (ii) In our investigation of the graph's propagation time, we discovered that a lesser number of layers (1 - 2 layers) results in inadequate information capture, while an increased number leads to the incorporation of noisy neighbors. As a result, we empirically select three layers for our experiments to achieve the best performance.

4.7 Computation Efficiency

The training time and latency of SaLa and SOTA methods (*i.e.*, Co-guiding Net, ReLa-Net and GISCo) are shown in Table 4. We find that our SaLa costs some more training time due to the maintenance of label co-occurrence matrices for various scenarios. Regarding latency, our SaLa is comparable to previous works, yet it significantly outperforms the latter. The graphs in our proposed SaLa only updated in the training process and are frozen during inference.

4.8 Single-intent and Multi-domain SLU

Since SaLa is scenario-aware, a natural question arises about its effectiveness in multi-domain SLU. To further assess the generalizability of our proposed SaLa framework, we conducted experiments on four extra benchmarks of both single-intent and multi-domain SLU settings. From the results reported in Tables 5 and Table 6, we can observe that the proposed SaLa not only achieves state-of-the-art performance across various SLU settings but also demonstrates more pronounced effectiveness in multi-domain SLU.

4.9 Case Study

To intuitively understand how the SaLa works, we provide two cases and visualize the label co-occurrence probabilities of the scenario of the utterance, as well as with three top confidence scores intents and one low confidence score intent. As shown in Figure 5(a), our SaLa can accurately predict multiple intent labels (*e.g.*, `AddToplaylist` and `PlayMusic`). Thanks to the group-level label co-occurrence matrices, the proposed SaLa can offer more precise guidance for interactions between labels, resulting in accurate prediction. Similar results can also be observed in Figure 5(b).

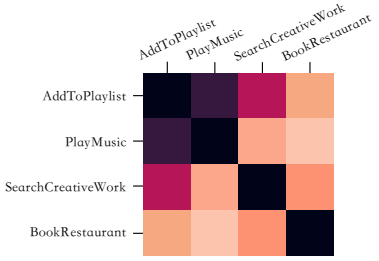
5 Related Work

Intent detection and slot filling are two subtasks in task-oriented dialogue systems that have become a research hotspot due to their ability to recognize and extract expressed intents and annotate corresponding sequence slot tags within a single utterance. Since intent detection and slot filling are highly correlated in SLU systems, numerous models [3, 4, 30, 40, 43, 44] have been proposed to jointly tackle the two subtasks. However, these models focus on single-intent utterances, which may not be practical in real-world scenarios where an utterance usually expresses multiple intents.

To this end, Kim et al. [20] begin to explore multiple intent detection. Gangadharaiah and Narayanaswamy [13] first employ a

(a)

Utterance	add	born	free	to	fresh	⋯	tune	from	the	twenties
Slot (Baseline)	O	B-entity_name	I-entity_name	O	B-playlist		O	O	O	B-year
Slot (Ours)	O	B-entity_name	I-entity_name	O	B-playlist		B-music_item	O	O	B-year
Intent (Baseline)	AddToPlaylist									
Intent (Ours)	AddToPlaylist, PlayMusic									



(b)

Utterance	what	⋯	canadian	airlines	⋯	flights	use	j31
Slot (Baseline)	O	⋯	B-airline_name	I-airline_name	⋯	O	O	B-aircraft_code
Slot (Ours)	O	⋯	B-airline_name	I-airline_name	⋯	O	O	B-aircraft_code
Intent (Baseline)	atis_quantity							
Intent (Ours)	atis_quantity, atis_city							

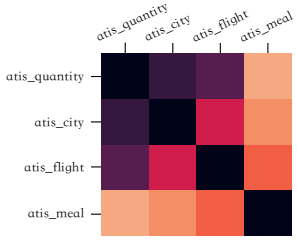


Figure 5: Case study of two different utterances between our SaLa and best baseline ReLa-Net on MixSNIPS (a) and MixATIS (b). Label in red denotes an error while the one in blue denotes the correct.

multi-task framework to tackle the multiple intent detection and slot filling jointly. With the increasing popularity of graph neural networks in various NLP tasks [5, 19, 42, 52], state-of-the-art multi-intent SLU systems also leverage Graph Neural Networks to model the cross-task interactions. Qin et al. [33] and Qin et al. [32] propose graph interaction networks to model implicit correlations between intent labels and slot tokens. Song et al. [36] build a global graph to leverage the intent-slot co-occurrence, enhancing the SLU performance. Xing and Tsang [46] implements a two-stage framework achieving mutual guidance between intents and slots. Xing and Tsang [47] further exploits label typologies and relations. Xing and Tsang [48] explores relational temporal graph reasoning to achieve fine-grained temporal modeling among labels.

Despite promising results achieved, current multi-intent SLU methods utilize labels at either the utterance level or the corpus level, failing to fully exploit an important feature of the SLU domain: scenario. This limitation hinders the progression towards more balanced and nuanced group-level label interactions.

6 Conclusion

In this paper, we proposed a novel scenario-aware label graph learning framework for multi-intent SLU. Concretely, the scenario-aware label co-occurrence module maintains a label co-occurrence matrix for each scenario and tracks co-occurring labels during the training phase, which is used to guide the interactions of label representations via graph propagation. Experimental results on two public benchmarks demonstrate the superiority of our SaLa framework. Future research will focus on addressing more complex SLU settings, like those involving unknown domains.

Acknowledgements

We thank all the anonymous reviewers for their insightful comment. This paper was partially supported by NSFC (No:62176008).

References

[1] Fengyu Cai, Wanhao Zhou, Fei Mi, and Boi Faltings. 2022. 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*. IEEE, 7607–7611. https://doi.org/10.1109/ICASSP43922.2022.9747477

[2] Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022. Towards joint intent detection and slot filling via higher-order attention. In *IJCAI*.

[3] Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022. Towards Joint Intent Detection and Slot Filling via Higher-order Attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 4072–4078. https://doi.org/10.24963/ijcai.2022/565

[4] Dongsheng Chen, Zhiqi Huang, and Yuexian Zou. 2022. Leveraging Bilinear Attention to Improve Spoken Language Understanding. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 7142–7146. https://doi.org/10.1109/ICASSP43922.2022.9747553

[5] Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced Multi-Channel Graph Convolutional Network for Aspect Sentiment Triplet Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2974–2985. https://doi.org/10.18653/v1/2022.acl-long.212

[6] Lisung Chen, Nuo Chen, Yuexian Zou, Yong Wang, and Xinzhong Sun. 2022. A Transformer-based Threshold-Free Framework for Multi-Intent NLU. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 7187–7192. https://aclanthology.org/2022.coling-1.629

[7] Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022. Joint Multiple Intent Detection and Slot Filling Via Self-Distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 7612–7616. https://doi.org/10.1109/ICASSP43922.2022.9747843

[8] Lizhi Cheng, Weijia Jia, and Wenmian Yang. 2021. An Effective Non-Autoregressive Model for Spoken Language Understanding. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, Gianluca Demartini,

Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 241–250. https://doi.org/10.1145/3459637.3482229

[9] Lizhi Cheng, Wenmian Yang, and Weijia Jia. 2022. A Scope Sensitive and Result Attentive Model for Multi-Intent Spoken Language Understanding. *CoRR* abs/2211.12220 (2022). https://doi.org/10.48550/arXiv.2211.12220 arXiv:2211.12220

[10] Xuxin Cheng, Zhihong Zhu, Bowen Cao, Qichen Ye, and Yuexian Zou. 2023. MRRL: Modifying the Reference via Reinforcement Learning for Non-Autoregressive Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10495–10505. https://doi.org/10.18653/v1/2023.findings-emnlp.704

[11] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *CoRR* abs/1805.10190 (2018). arXiv:1805.10190 http://arxiv.org/abs/1805.10190

[12] 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*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 5467–5471. https://doi.org/10.18653/v1/p19-1544

[13] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 564–569. https://doi.org/10.18653/v1/n19-1055

[14] Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 753–757. https://doi.org/10.18653/v1/n18-2118

[15] Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm.. In *Interspeech*. 715–719.

[16] Mutian He and Philip N Garner. 2023. Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding. *arXiv preprint arXiv:2305.13512* (2023).

[17] Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[19] Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 4820–4829. https://doi.org/10.18653/v1/D19-1488

[20] Byeongchang Kim, Seonghan Ryu, and Gary Geunbae Lee. 2017. Two-stage multi-intent detection for spoken language understanding. *Multim. Tools Appl.* 76, 9 (2017), 11377–11390. https://doi.org/10.1007/s11042-016-3724-4

[21] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard Product for Low-rank Bilinear Pooling. In *International Conference on Learning Representations*.

[22] Young-Bum Kim, Sungjin Lee, and Karl Stratos. 2017. Onenet: Joint domain, intent, slot prediction for spoken language understanding. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 547–553.

[23] Jihwan Lee, Ruhi Sarikaya, and Young-Bum Kim. 2019. Locale-agnostic Universal Domain Classification Model in Spoken Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. 9–15.

[24] Bing Liu and Ian Lane. 2017. Multi-domain adversarial learning for slot filling in spoken language understanding. *arXiv preprint arXiv:1711.11310* (2017).

[25] 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 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[26] Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A Coarse-to-Fine Approach for Cross-domain Slot Filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 19–25.

[27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[28] OpenAI. 2023. https://chat.openai.com/. 2023.

[29] 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 2078–2087. https://doi.org/10.18653/v1/D19-1214

[30] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8193–8197.

[31] Libo Qin, Fuxuan Wei, Minheng Ni, Yue Zhang, Wanxiang Che, Yangming Li, and Ting Liu. 2022. Multi-domain spoken language understanding using domain-and task-aware parameterization. *Transactions on Asian and Low-Resource Language Information Processing* 21, 4 (2022), 1–17.

[32] 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 178–188. https://doi.org/10.18653/v1/2021.acl-long.15

[33] Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020. Towards Fine-Grained Transfer: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1807–1816. https://doi.org/10.18653/v1/2020.findings-emnlp.163

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[35] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3795–3805.

[36] Mengxiao Song, Bowen Yu, Quangang Li, Yubin Wang, 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 7967–7977. https://aclanthology.org/2022.emnlp-main.543

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

[38] 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*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[39] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph Attention Networks. *CoRR* abs/1710.10903 (2017). arXiv:1710.10903 http://arxiv.org/abs/1710.10903

[40] Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2021. Encoding Syntactic Knowledge in Transformer Encoder for Intent Detection and Slot Filling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13943–13951. https://ojs.aaai.org/index.php/AAAI/article/view/17642

[41] Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 309–314. https://doi.org/10.18653/v1/n18-2050

[42] Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual Text Classification with Heterogeneous Graph Neural Network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 612–620. https://doi.org/10.18653/v1/2021.acl-short.78

[43] 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 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1932–1937. https://doi.org/10.18653/v1/2020.emnlp-main.152

[44] Jie Wu, Ian G. Harris, and Hongzhi Zhao. 2021. Spoken Language Understanding for Task-oriented Dialogue Systems with Augmented Memory Networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 797–806. https://doi.org/10.18653/v1/2021.naacl-main.63

[45] 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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 7203–7208. https://aclanthology.org/2022.coling-1.631

[46] Bowen Xing and Ivor W. Tsang. 2022. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 159–169. https://aclanthology.org/2022.emnlp-main.12

[47] Bowen Xing and Ivor W. Tsang. 2022. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 3964–3975. https://aclanthology.org/2022.emnlp-main.263

[48] Bowen Xing and Ivor W Tsang. 2023. Relational Temporal Graph Reasoning for Dual-task Dialogue Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[49] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.

[50] Linhao Zhang, Dehong Ma, Xiaodong Zhang, Xiaohui Yan, and Houfeng Wang. 2020. Graph lstm with context-gated mechanism for spoken language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 9539–9546.

[51] Xiaodong Zhang and Houfeng Wang. 2016. 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, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 2993–2999. http://www.ijcai.org/Abstract/16/425

[52] Zheng Zhang, Zili Zhou, and Yanna Wang. 2022. SSEGCN: Syntactic and Semantic Enhanced Graph Convolutional Network for Aspect-based Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 4916–4925. https://doi.org/10.18653/v1/2022.naacl-main.362

[53] Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Enhancing Code-Switching for Cross-lingual SLU: A Unified View of Semantic and Grammatical Coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7849–7856. https://aclanthology.org/2023.emnlp-main.486

[54] Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023. Towards Unified Spoken Language Understanding Decoding via Label-aware Compact Linguistics Representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12523–12531. https://doi.org/10.18653/v1/2023.findings-acl.793

[55] Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou. 2022. A Dynamic Graph Interactive Framework with Label-Semantic Injection for Spoken Language Understanding. *CoRR* abs/2211.04023 (2022). https://doi.org/10.48550/arXiv.2211.04023 arXiv:2211.04023