Towards Explainable Joint Models via Information Theory for Multiple Intent Detection and Slot Filling

Xianwei Zhuang*, **Xuxin Cheng***, **Yuexian Zou**†

School of ECE, Peking University, China
{xwzhuang, chengxx}@stu.pku.edu.cn, zouyx@pku.edu.cn

Abstract

Recent joint models for multi-intent detection and slot filling have obtained promising results through modeling the unidirectional or bidirectional guidance between intent and slot. However, existing works design joint models heuristically and lack some theoretical exploration, including (1) theoretical measurement of the joint-interaction quality; (2) explainability of design and optimization methods of joint models, which may limit the performance and efficiency of designs. In this paper, we mathematically define the cross-task information gain (**CIG**) to measure the quality of joint processes from an information-theoretic perspective and discover an implicit optimization of CIG in previous models. Based on this, we propose a novel multi-stage iterative framework with theoretical effectiveness, explainability, and convergence, which can explicitly optimize information for cross-task interactions. Further, we devise an **info**rmation-based **joint** model (**InfoJoint**) that conforms to this theoretical framework to gradually reduce the cross-task propagation of erroneous semantics through CIG iterative maximization. Extensive experiment results on two public datasets show that InfoJoint outperforms the state-of-the-art models by a large margin.

Introduction

Spoken language understanding (SLU) is a critical task in dialog systems (Young et al. 2013), which generally includes two subtasks: intent detection (ID) and slot filling (SF) (Tur and De Mori 2011). Recently, joint models (Goo et al. 2018; Qin et al. 2019; Cheng et al. 2023c; Zhu et al. 2023a,b) for ID and SF have achieved impressive performance, and have proved that there exists a strong correlation between this two tasks (Weld et al. 2022). The recent studies (Gangadharaiah and Narayanaswamy 2019) recognize that a single utterance often contains multiple intentions in real-world scenarios, *i.e.*, Multi-Intent SLU. Thus, joint multi-intent SLU gradually attracting increasing attention. One of the mainstream insights works on devising attention mechanisms against joint tasks. AGIF (Qin et al. 2020, 2021b) explore unidirectional joint models from ID to SF based on graph attention (GAT) for multi-intent SLU. Cheng, Yang, and Jia (2023) proposes a variant of attention for reducing error propagation. Co-guiding Net (Xing and Tsang 2022) proposes heterogeneous attention (HGAT) to achieve bidirectional interaction. These works heuristically design cross-task interaction modules to improve performance.

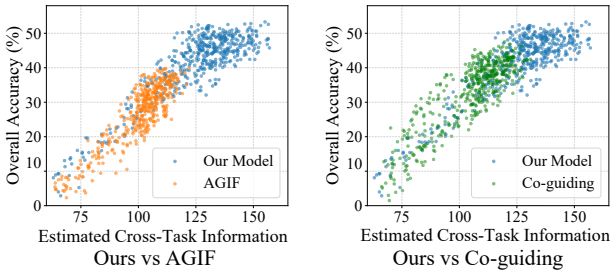


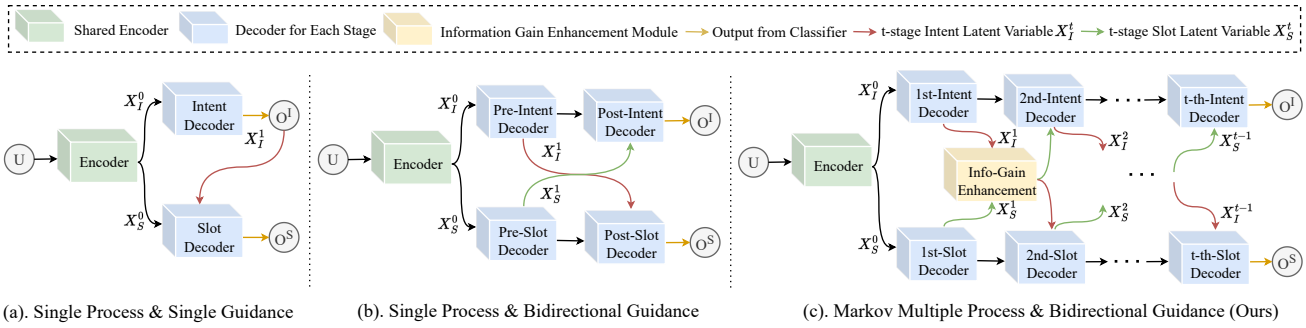
Figure 1: Comparison of statistical results based on the perspective of cross-task information (in color). We run each model five times on the MixATIS (Qin et al. 2020) to obtain statistical results, and observe that: (i) There is an implicit positive correlation between the performance of prior joint models and the CIG. (ii) Our joint model explicitly optimizes cross-task information to obtain higher interactive information and better performance.

Although promising progress has been made in previous works, these studies regrettably lack exploration of some core issues at the theoretical level, including:

(1) Why does the joint SLU model work better and how to quantitatively measure the enhancement quality of joint models in theory? Our work innovatively defines the cross-information gain (CIG) to measure the interaction quality of the beneficial information between ID and SF, and explains that the process of joint optimization is the process of **maximizing CIG** and **minimizing the information gap** between dual-task branches. Based on this perspective, we discover that previous works implicitly improve CIG although they do not explicitly mention CIG (*i.e.*, there is a positive correlation between CIG and performance, as shown in Fig. 1).

(2) How to design an explainable model for the joint processes? Existing joint models are essentially single-stage models (Fig. 2a and b) which perform single interaction, and lack theoretical guidance. We construct the model as a

*These authors contributed equally.
†Corresponding author.



(a). Single Process & Single Guidance (b). Single Process & Bidirectional Guidance (c). Markov Multiple Process & Bidirectional Guidance (Ours)

Figure 2: Comparison of different joint frameworks for multi-intent SLU at the abstract level. (a) The framework of vanilla joint models. The model only models the unidirectional guidance from multi-intent detection to SF in a single process. (b) The framework considers mutual guidance between the two tasks, but only heuristically designs various modules in a single stage to achieve dual-task interaction. (c) Our framework models the dual-task interaction process as a Markov process and ensures the effectiveness, theoretical interpretability, and convergence of bidirectional guidance from an information theoretic perspective. Our framework implements iterative enhancement of the dynamic interaction for multi-intent SLU.

multi-stage Markov model (Fig. 2c) and claim the method can iteratively optimize in the positive direction and outperform single-stage models under some theoretical constraints. We further analyze the theoretical constraints for its effectiveness and convergence.

Furthermore, we devise a novel information-based joint model (InfoJoint) which satisfies the above theoretical constraints. Specifically, InfoJoint utilizes multi-level cross-task contrastive learning to maximize CIG. In addition, directional constraints and entropy weighting strategies are developed to facilitate positive optimization and balance learning in each batch, respectively. Finally, InfoJoint optimizes CIG through iterative training to achieve deep interaction and gradually eliminate error propagation between the joint tasks. In summary, our framework has several appealing facets: (1) **Explainability:** Joint models are guided by theory to fully explore relevance through iterative enhancement. (2) **Convergence:** The convergence of joint optimization is theoretically guaranteed. (3) **Universality:** This framework is architecture-independent (component-independent) and compatible with previous works. The main contributions of this paper are presented as follows:

(1) **(Theory)** We propose a novel explainable multi-stage joint framework for multi-intent SLU with theoretical quantifiability, effectiveness, and convergence. (2) **(Methodology)** Based on the information-based principled framework, we devise an iterative-enhancement model termed InfoJoint, which adopts multi-level cross-task contrastive learning, directional constraints, and entropy weighting to achieve the effective interaction and performance improvement. (3) **(Experiments)** Extensive experiments on two public multi-intent SLU datasets MixATIS and MixSNIPS demonstrate that InfoJoint significantly outperforms the best baselines in terms of all evaluation metrics.

Related Work

Joint model for intent detection and slot filling. Early studies (Yao et al. 2014; Kurata et al. 2016; Zhang and Wang 2016) recognize that there is a close connection between ID and SF. Motivated by this, some studies (Goo et al. 2018; E et al. 2019; Liu et al. 2019a; Qin et al. 2019; Zhang et al. 2019; Wu et al. 2020; Qin et al. 2021a; Ni et al. 2023) start to model the relationship between ID and SF in a multi-task manner to improve the performance. However, these works ignore the multi-intent context in real-world scenarios, which are more challenging.

Joint model for multi-intent SLU. Kim, Ryu, and Lee (2017) start to explore multi-intent scene recognition and Gangadharaiah and Narayanaswamy (2019) propose the first joint model for multi-intent SLU. Qin et al. (2020, 2021b) propose GAT-based joint models to improve interaction performance, and Xing and Tsang (2022) propose a HGAT-based model to achieve the bidirectional interaction. GISCO (Song et al. 2022) considers the global correlation between ID and SF. SSRAN (Cheng, Yang, and Jia 2023) proposes a scope-sensitive attention network to model the dual-task interaction. Different from the above methods, we emphasize maximizing CIG iteratively to achieve the explainable cross-task interaction, which gives a significant insight on improving the performance of multi-intent SLU.

Theoretical Analysis

Problem Definition

We design a joint-learning network $\Psi(\cdot;\theta)$ that directly consumes a input utterance $U = \{u_i\}_{i=1}^n$ for multi-intent SLU. The model obtains predictions for multi-label intent classification $O^I = \{o_i^I\}_{i=1}^m$ and slot labels $O^S = \{o_i^S\}_{i=1}^n$ that map the utterance U, where m denotes the number of intents in a given utterance and n denotes the utterance length. We further define the training set of utterances as $\mathcal{D}^U = \{U\}$.

Theoretical Formulation

Markov Modeling. As shown in Fig. 2a and Fig. 2b, to go beyond the single-stage heuristic modeling of dual-task interaction, we model the joint enhancement process as a multi-stage Markov process which is illustrated in Fig. 2c. This means that the hidden state random variables in the next stage $t+1$ are only related to the current state t for $t \in \mathcal{N}$.

Formally, we obtain the context-sensitive hidden states $X_I^0 = \{x_{[I,i]}^0\}_{i=1}^n$ and $X_S^0 = \{x_{[S,i]}^0\}_{i=1}^n$ through the shared encoder, where I and S represent ID and SF respectively. Note that the two tasks share the same semantics features in the initial stage, *i.e.*, $X_I^0 = X_S^0$. We further define the task-specific features obtained in the t-th iteration stage of the intent-detection branch and slot-filling branch as random variables $X_I^t = \{x_{[I,i]}^t\}_{i=1}^n$ and $X_S^t = \{x_{[S,i]}^t\}_{i=1}^n$.

Cross-Task Information Gain. To explicitly express the relationship between the two tasks, we consider quantifying the degree of joint enhancement from an information theoretic perspective, and mathematically defining the CIG of the two tasks (implicit random variable X_i and X_j):

$$\mathcal{CIG}(X_i; X_j) = \mathcal{H}_\theta(X_i) - \mathcal{H}_\theta(X_i \mid X_j), \tag{1}$$

where $\mathcal{H}_\theta(\cdot)$ denotes information entropy under the model θ, and $i = I, j = S$ or $i = S, j = I$. CIG is mathematically symmetric and non-negative, which shows a reduction of the uncertainty of the variable (task) X_i given another variable (task) X_j. Based on this, the assumption that the two tasks X_i and X_j are mutually reinforcing can be expressed as:

$$\mathcal{CIG}(X_i; X_j) = \mathcal{CIG}(X_j; X_i) > 0. \tag{2}$$

From this perspective, we explain that the designs of previous joint models essentially strive to improve CIG to obtain task-specific bidirectional information and thereby improve prediction performance.

Optimization Constraints. Our model exploits CIG to improve performance on the premise that the following three constraints are satisfied during the optimization process:

Proposition 1 *(Intrinsic Constraint) Suppose t is the number of steps for the iterative optimization process, $\forall t \in \mathcal{N}^+$,*

$$\mathcal{CIG}(X_i^t; X_j^{t-1}) > 0, \tag{3}$$

where $i = I, j = S$ or $i = S, j = I$.

Proposition 1 is the foundation of the joint model, which mathematically describes that the amount of beneficial information provided by one task to another is always positive.

Proposition 2 *(Directional Constraint) Suppose t_x and t_y are the number of steps for the iterative optimization process, $\forall t_x, t_y \in \mathcal{N}^+$, if $t_x > t_y$, then,*

$$\mathcal{CIG}(X_i^{t_x}; X_j^{t_x-1}) > \mathcal{CIG}(X_i^{t_y}; X_j^{t_y-1}), \tag{4}$$

where $i = I, j = S$ or $i = S, j = I$.

Proposition 2 constrains the function $\mathcal{CIG}(X_i^t; X_j^{t-1})$ to monotonically increase *w.r.t.* the number of iterations t. Intuitively, a better expression of the current task will provide more information gain for another. This proposition ensures that the model is positively optimized in each iteration.

Proposition 3 *(Upper-bound Constraint) The non-continuous set function $\mathcal{CIG}(X_i^t; X_j^{t-1})$ w.r.t the number of iterative steps t has a supremum $\mathcal{MI}(O^I, O^S)$, i.e.,*

$$\sup\ \{\mathcal{CIG}(X_i^t; X_j^{t-1}) : t \in \mathcal{N}^+\} = \mathcal{MI}(O^I, O^S), \tag{5}$$

where $\mathcal{MI}(O_I, O_S)$ is defined as the mutual information between intent-detection labels O^I and slot-filling labels O^S, and $i = I, j = S$ or $i = S, j = I$.

Proposition 3 is a boundary condition that describes the existence of an upper bound on the information gain provided by the two tasks to each other. Further, this upper bound is determined by the optimal predictions ($i.e.$, labels O^I and O^S) of the two tasks.

Convergence Analysis. We next show the convergence ability of our joint model for universal optimization during the iteration process. By the convergence of functions, intuitively, a model should approach the limit value, *i.e.*, the optimal value, after sufficient limited optimization steps.

Theorem 1 *(Convergence Property) Suppose $\mathcal{T} = \{t : t \in \mathcal{N}^+\}$, $\mathcal{CIG} : \mathcal{T} \to \mathbb{R}^+$ is a non-continuous set function w.r.t the number of iterative steps t and simultaneously satisfies Proposition 1-3, then, $\forall \epsilon > 0, \exists \delta > 0$, for $t \in \mathcal{T}$, if $t > \delta$,*

$$\mid \mathcal{CIG}(X_i^t; X_j^{t-1}) - \mathcal{MI}(O^I; O^S) \mid < \epsilon. \tag{6}$$

Proof. Through Proposition 1-3, we can derive that the non-continuous set function $\mathcal{CIG}(X_i^t; X_j^{t-1})$ is a monotonically increasing bounded function. According to Monotone Convergence Theory (Yeh 2006), it can be further derived that the function converges to the supremum $\mathcal{MI}(O^I; O^S)$.

Theorem 6 theoretically demonstrates the convergence of our framework. To further measure convergence, the following energy function $\mathbf{E}(\cdot)$ can be defined to estimate the iteration situation as $\mathbf{E}(t) = \frac{\mathcal{CIG}(X_i^t; X_j^t)}{\mathcal{MI}(O_I; O_S)}$. We note that there is no explicit analytical expression of $\mathcal{CIG}$ about t, and therefore the energy function $\mathbf{E}(t)$ cannot be directly calculated. Following previous works (Bugliarello et al. 2020; Ji et al. 2022a), we introduce a Monte Carlo estimator in the training set $\mathcal{D}^U$ to approximate $\mathbf{E}(t)$.

Method

Architecture Overview

In this section, we describe the overview of InfoJoint as shown in Fig. 3a and introduce the relationship between theory and design. **Firstly**, to avoid overparameterization caused by a large number of decoders, we encode the time t and then concatenate it with the input embedding of decoders to achieve parameter reuse. **Secondly**, X_i^t and X_j^{t-1} are obtained asynchronously. We realize that $\mathcal{CIG}(X_i^t; X_j^{t-1}) \to \mathcal{CIG}(X_i^t; X_j^t)$ when the model is sufficiently iterated. Therefore, we utilize $\mathcal{CIG}(X_i^t; X_j^t)$ to approximate $\mathcal{CIG}(X_i^t; X_j^{t-1})$ to achieve the synchronous optimization. **Thirdly**, inspired by Ji et al. (2022b); Cheng et al. (2023a), we utilize contrastive learning to minimize the **InfoNCE** loss to maximize a lower bound on CIG.

Following the above rules, the general framework of our InfoJoint is shown in Fig. 3, which consists of four core components: a shared encoder, a time encoding module, a multi-level cross-task contrastive learning (**MCCL**) module and bidirectional information extraction modules. We utilize MCCL and bidirectional information extraction modules to gradually extract and improve the CIG through t iterations. Instead of treating all input utterances indiscriminately, we propose an entropy-based weighting strategy to balance utterance information in each batch.

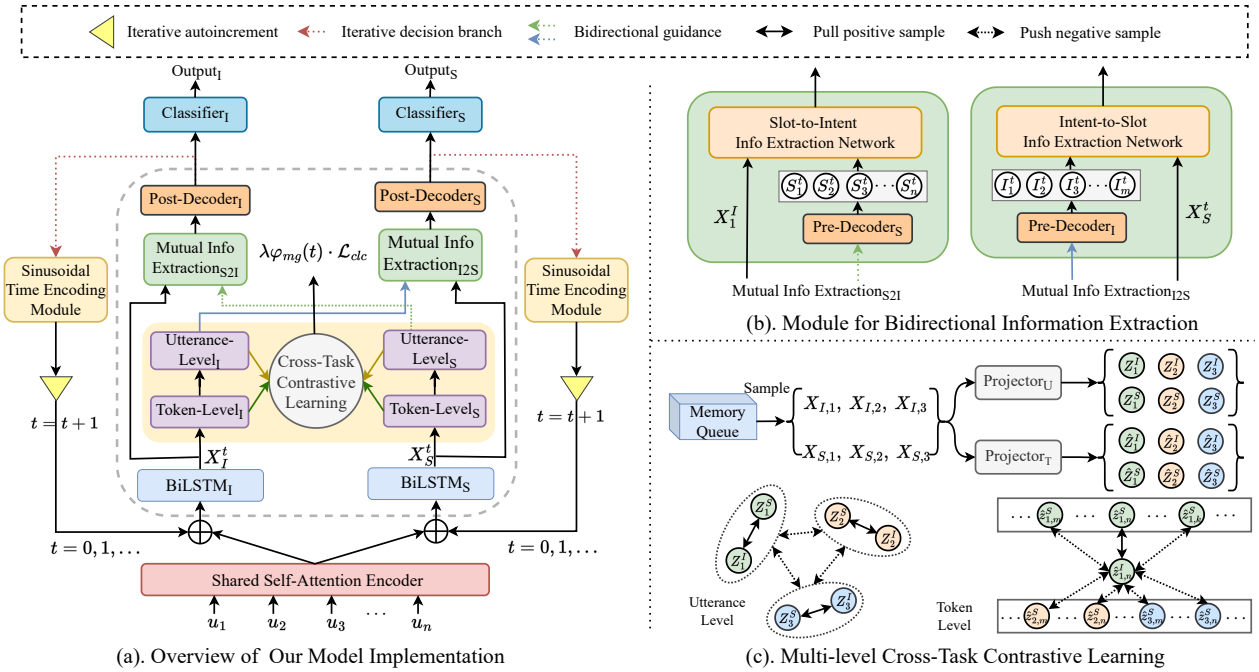


Figure 3: A engineering implementation of our theoretical framework. (a) The overall architecture of our proposed InfoJoint based on information theory. (b) We demonstrate the module for extracting bidirectional information in our joint model. (c) We show the multi-level cross-task contrastive learning strategy in each iteration stage to gradually improve CIG.

Shared Encoder and Time Encoding

Following Qin et al. (2020, 2021b); Xing and Tsang (2022), we adopt a bidirectional LSTM (BiLSTM) to produce a series of context-sensitive hidden states $\mathbf{H} = \{h_i\}_{i=1}^n$ over the word embeddings $\hat{U} = \{\hat{u}_i\}_{i=1}^n$, and a self-attention mechanism to capture context-aware features $\mathbf{A} \in \mathbb{R}^{n \times d_k}$:

$$h_i = \text{BiLSTM}(\hat{u}_i, h_{i-1}, h_{i+1}),$$
$$\mathbf{A} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}, \tag{7}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are matrices obtained by mapping the input word vector through different linear projections, and d_k denotes the dimension of keys $\mathbf{K}$.

We finally fuse these two representations as the encoding features: $\mathbf{E}_I^0 = \mathbf{E}_S^0 = \mathbf{H}||\mathbf{A}$, where $||$ is a concatenation operation. To distinguish the decoder states at different stages, we utilize sine position encoding (Vaswani et al. 2017) to encode the iteration time t into a time encoding vector $\mathbf{T} \in \mathbb{R}^{n \times 2d_k}$. Then, we sum the basic semantics (*i.e.*, $\mathbf{E}_I^{t-1}$ or $\mathbf{E}_S^{t-1}$) and time encoding $\mathbf{T} = \{T_i\}_{i=1}^n$ as the features of stage t to participate in iterative optimization, where $t > 1$ indicates that the semantic feature $\mathbf{E}$ is from the previous stage instead of the encoder. The iteration time t will auto-increment during optimization before it exceeds the maximum time T_{max}. Based on this, we further obtain task-specific features $X_k^t = \{x_{[k,i]}^t\}_{i=1}^n$ for multi-intent detection I and slot filling S through a layer of BiLSTM:

$$x_{[k,i]}^t = \text{BiLSTM}_\text{k}(E_{[k,i]}^{t-1} + T_i, x_{[k,i-1]}^t, x_{[k,i+1]}^t), \tag{8}$$

where $k = I$ and $k = S$ represent ID and SF respectively.

Bidirectional Information Extraction

These two modules (shown in Fig. 3b) are mainly proposed to extract the valuable information which is conducive to obtaining better expression from the two tasks, namely extracting $\mathcal{CIG}(X_i^t; X_i^{t-1}, X_j^{t-1})$. We employ the Heterogeneous Graph Attention Network (HGAT) (Velickovic et al. 2018; Wang et al. 2019) to implement this bidirectional guidance.

Specifically, to explicitly leverage slot information to guide multi-intent detection, we obtain the estimated slot labels sequence $S^t = \{S_i^t\}_{i=1}^n$ by a pre-decoder, and construct a slot-to-intent semantic graph $\mathcal{G}_{S2I} = (\mathcal{V}_{S2I}, \mathcal{E}_{S2I})$ similar to Xing and Tsang (2022). Further, we feed the node semantics $H_I^t = \{h_{[I,i]}^t\}_{i=1}^n = \{x_i^t, \phi^{emb}(S_i^t)\}_{i=1}^n$ into HGAT with K attention heads and finally produce more effective representation $\mathbf{E}_I^t = \{E_{[I,i]}^t\}_{i=1}^n$:

$$\mathcal{F}\left(h_{[I,i]}^t, h_{[I,j]}^t\right) = \mathbf{a}^{[k,r]^\top}\left[W_I^{[k,r]}h_{[I,i]}^t \| W_I^{[k,r]}h_{[I,j]}^t\right], \tag{9}$$

$$\alpha_{ij}^{[k,r]} = \frac{\exp\left(\sigma\left(\mathcal{F}\left(h_{[I,i]}^t, h_{[I,j]}^t\right)\right)\right)}{\sum_{j' \in \mathcal{N}_i}\exp\left(\sigma\left(\mathcal{F}\left(h_{[I,i]}^t, h_{[I,j']}^t\right)\right)\right)},$$
$$E_{[I,i]}^t = \Big\|_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}_i}\alpha_{ij}^{[k,r]}W_I^{[k,r]}h_{[I,j]}^t\right), \tag{10}$$

where $\phi^{emb}(\cdot)$ is a projector; r denotes the type of edge from node j to node i; σ represents the activation function; set

$\mathcal{N}_i$ is the first-order neighbors of node i on graph $\mathcal{G}_{S2I}$; and $\mathbf{a}^{[k,r]}$ and $W_I^{[k,r]}$ are trainable matrix of r on the k-th head.

We can adopt a similar approach to extract intent-to-slot information and obtain a more abstract feature $\mathbf{E}_S^t$ for Eq. 8.

Maximize Cross-Task Information Gain

Based on the existing theoretical discovery (Poole et al. 2019) that InfoNCE loss (Oord, Li, and Vinyals 2018) lower bound the mutual information, we propose MCCL (as shown in Fig. 3c) to estimate bidirectional information (*i.e.*, CIG) in practice. Specifically, we adopt a memory bank strategy (He et al. 2020) to obtain more effective negative samples. At each iteration, the model samples a set of matching pairs $\mathcal{M}^X = \{(X_{I,\tilde{k}}, X_{S,\tilde{k}})\}_{\tilde{k}=1}^{|\mathcal{M}^X|}$ in the memory queue. Subsequently, we perform contrastive learning at the levels of utterance and token to fully estimate and optimize CIG:

Utterance level. We map the matching sample set $\mathcal{M}^X$ to the d_U-dimension space and obtain a matching set in embedding space $\mathcal{S}^U = \{(Z_{\tilde{k}}^I, Z_{\tilde{k}}^S)\}_{\tilde{k}=1}^{|\mathcal{S}^U|}$ through the utterance-level projector, where $Z_{\tilde{k}}^I, Z_{\tilde{k}}^S \in \mathbb{R}^{1\times d_U}$. The **utterance-level contrastive loss** $\mathcal{L}_U$ can be formulated as:

$$\mathcal{L}_U = -\frac{1}{|\mathcal{S}^U|}\sum_{\tilde{k}=1}^{|\mathcal{S}^U|}\log\frac{\exp(Z_{\tilde{k}}^I \cdot Z_{\tilde{k}}^S/\tau_1)}{\sum_{j=1,j\neq\tilde{k}}^{|\mathcal{S}^U|}\exp(Z_{\tilde{k}}^I \cdot Z_j^S/\tau_1)}, \quad (11)$$

where τ_1 denotes the temperature coefficient.

Token level. We map $\mathcal{M}^X$ to the d_T-dimension space and obtain $\tilde{K}$ embeddings $\hat{Z}_{\tilde{k}}^I = \{\hat{z}_{\tilde{k},i}^I\}_{i=1}^n$ and $\hat{Z}_{\tilde{k}}^S = \{\hat{z}_{\tilde{k},i}^S\}_{i=1}^n$ through the projector, where $\hat{z}_{\tilde{k},i}^I, \hat{z}_{\tilde{k},i}^S \in \mathbb{R}^{1\times d_T}$. We further integrate the tokens of all utterances to form a token-level matching set $\mathcal{S}^T = \{(\hat{z}_{k'}^I, \tilde{z}_{k'}^S)\}_{k'=1}^{\tilde{K}\times n}$. The **token-level contrastive loss** $\mathcal{L}_T$ can be formulated as:

$$\mathcal{L}_T = -\frac{1}{|\mathcal{S}^T|}\sum_{k'=1}^{|\mathcal{S}^T|}\log\frac{\exp(\hat{z}_{k'}^I \cdot \hat{z}_{k'}^S/\tau_2)}{\sum_{j=1,j\neq k'}^{|\mathcal{S}^T|}\exp(\hat{z}_{k'}^I \cdot \hat{z}_j^S/\tau_2)}, \quad (12)$$

where τ_2 denotes the temperature coefficient.

Training and Inference

Entropy weighting. We find that the information of utterances trained synchronously in a batch is imbalanced. To alleviate this issue, we propose a weighting strategy based on information entropy. Specifically, we define semantic uncertainty $\tilde{\mathcal{H}}$ as the indicator of utterance importance, and further obtain the weight α^U of the utterance U in the batch $\mathcal{B}$:

$$\tilde{\mathcal{H}}(U) = -\sum_{i=1}^{n} p_i^u \log p_i^u, \quad \alpha^U = \frac{\tilde{\mathcal{H}}(U)}{\sum_{U\in\mathcal{B}}\tilde{\mathcal{H}}(U)}, \quad (13)$$

where p_i^u denotes the statistic frequency of i-th word u_i in the U under the training set $\mathcal{D}^U$.

Directional constraints. To further ensure each iteration of our InfoJoint model is positively optimized (*i.e.*, satisfying Proposition 2), we define an Exponential Decay Function $\varphi_{mg}(t)$ as a margin penalty for MCCL method, where, $\varphi_{mg}(t) = a \cdot e^{-kt}$, $a = 0.9$ and $k = 2$. Therefore, the total contrastive loss for stage t is:

$$\mathcal{L}_{mccl}^t = \varphi_{mg}(t) \cdot \mathcal{L}_{clc} = \varphi_{mg}(t) \cdot (\mathcal{L}_U + \mathcal{L}_T). \quad (14)$$

For stage t, $\mathbf{E}_I^t$ and $\mathbf{E}_S^t$ are fed to intent and slot post-decoder, producing the intent and slot label distributions for each utterance: $\mathbf{Y}_I^t$ and $\mathbf{Y}_S^t$. Then, the loss of intent $\mathbf{Y}_I^t$ and slot $\mathbf{Y}_S^t$ predictions can be calculated by the binary cross-entropy loss and negative log-likelihood loss: $\mathcal{L}_I^t$ and $\mathcal{L}_S^t$. Intuitively, the predictions in the t stage should be better than those in the $t-1$ stage. We further design another margin penalty $\mathcal{L}_{mg}^t$ for this rule ($i.e.$, Proposition 2):

$$\mathcal{L}_{mg}^t = \max\{0, \mathcal{L}_I^{t-1} - \mathcal{L}_I^t\} + \max\{0, \mathcal{L}_S^{t-1} - \mathcal{L}_S^t\}. \quad (15)$$

Therefore, the joint loss function of InfoJoint is:

$$\mathcal{L}^t = \alpha\mathcal{L}_I^t + (1-\alpha)\mathcal{L}_S^t + \beta\mathcal{L}_{mg}^t + \lambda\mathcal{L}_{mccl}^t, \quad (16)$$

where α, β and λ are trade-off hyper-parameters.

Details. In the training stage, InfoJoint adopts Eq. 8 to perform iterative optimization from $t = 1$ to $t = T_{max}$ for each batch. For efficiency, we manually define T_{max} instead of using the Energy function to determine convergence. In the inference stage, the iteration process and MCCL can be discarded without affecting the inference speed. We can directly obtain inference results by setting t as the optimal T_{max}. And the final outputs O^I and O^S are obtained via applying the Top-K strategy over $\mathbf{Y}_I^t$ and $\arg\max$ over $\mathbf{Y}_S^t$.

Experiments

Experimental Settings

Datasets and Metrics. Following previous works, we conduct our experiments on two public multi-intent SLU datasets[1] to evaluate the effectiveness of InfoJoint, *i.e.*, the cleaned version of MixATIS and MixSNIPS (Hemphill, Godfrey, and Doddington 1990; Coucke et al. 2018; Qin et al. 2020). MixATIS and MixSNIPS include 13162, 759, 828 utterances and 39776, 2198, 2199 ones for training, validation and testing respectively. For a fair comparison with previous works, we also adopt accuracy(Acc), F1 score and overall accuracy as metrics for multi-intent detection, SF and sentence-level semantic frame parsing, respectively.

Settings. The hyper-parameters α, β and λ of loss (Eq. 16) are set as 0.7, 0.2 and 0.4 on MixATIS, and 0.6, 0.2 and 0.4 on MixSNIPS. We adopt grid search to determine hyperparameters for optimal performance. The temperature τ_1 and τ_2 in Eq. 11 and 12 are empirically set as 0.05. We utilize Adam (Kingma and Ba 2015) with a learning rate of 0.001 and a weight decay of $1e^{-6}$ to train InfoJoint for both datasets. We train all models from scratch with 100 epochs. For batch size, we set 16 and 32 for MixATIS and MixSNIPS. All experiments are conducted on 4 RTX3090 GPUs.

[1] https://github.com/LooperXX/AGIF

Model	MixATIS			MixSNIP		
	Intent	Slot	Overall	Intent	Slot	Overall
Bi-Model (Wang, Shen, and Jin 2018)	70.3	83.9	34.4	95.6	90.7	63.4
Stack-Propagation (Qin et al. 2019)	72.1	87.8	40.1	96.0	94.2	72.9
Joint Multiple ID-SF (Gangadharaiah and Narayanaswamy 2019)	73.4	84.6	36.1	95.1	90.6	62.9
AGIF (Qin et al. 2020)	74.4	86.7	40.8	95.1	94.2	74.2
GL-GIN (Qin et al. 2021b)	76.3	88.3	43.5	95.6	94.9	75.4
GISCO (Song et al. 2022)	75.0	88.5	48.2	95.5	95.0	75.9
Co-guiding Net (Xing and Tsang 2022)	79.1	89.8	51.3	97.7	95.1	77.5
SSRAN (Cheng, Yang, and Jia 2023)	77.9	89.4	48.9	<u>98.4</u>	<u>95.8</u>	77.5
ChatGPT (OpenAI 2023)	66.1	43.7	34.2	94.9	59.4	39.6
InfoJoint(T_{max}=5)	<u>79.8</u>	<u>90.6</u>	<u>51.9</u>	97.9	**96.9***	<u>78.2</u>
InfoJoint(T_{max}=10)	**80.6***	**91.4***	**52.5***	**99.2***	**96.9***	**78.9***

Table 1: Quantitative comparison results on MixATIS and MixSNIPS. * denotes the improvement of InfoJoint over all baselines is statistically significant with $p < 0.05$ under t-test. The best results are in bold and the second best ones are underlined.

	Multi-level Contrastive Loss	Directional constraints.	Time Coding	Entropy Weighting	MixATIS			MixSNIPS		
					Intent	Slot	Overall	Intent	Slot	Overall
(a)	×	×	×	×	74.7	88.2	43.2	96.1	94.4	74.5
(b)	$\mathcal{L}_T$	$\varphi_{mg}(t) + \mathcal{L}_{mg}$	Sinusoidal	✓	76.8	88.5	46.2	96.8	94.6	76.4
(c)	$\mathcal{L}_U$	$\varphi_{mg}(t) + \mathcal{L}_{mg}$	Sinusoidal	✓	78.6	89.5	49.7	97.6	95.2	77.5
(d)	$\mathcal{L}_U + \mathcal{L}_T$	×	Sinusoidal	✓	77.5	88.7	47.6	96.8	95.3	76.9
(e)	$\mathcal{L}_U + \mathcal{L}_T$	$\mathcal{L}_{mg}$	Sinusoidal	✓	77.8	88.9	48.4	97.3	95.6	77.3
(f)	$\mathcal{L}_U + \mathcal{L}_T$	$\varphi_{mg}(t) + \mathcal{L}_{mg}$	Sinusoidal	×	79.7	**91.5**	51.8	98.4	96.5	78.2
(g)	$\mathcal{L}_U + \mathcal{L}_T$	$\varphi_{mg}(t) + \mathcal{L}_{mg}$	Sinusoidal	✓	**80.6**	91.4	**52.5**	**99.2**	**96.9**	**78.9**

Table 2: Ablation study on both datasets for quantitatively evaluating the contribution of different components to InfoJoint. We repeat this experiment five times to obtain the statistical mean.

Main Results and Analysis

The main experimental results are shown in Table 1. We can see that InfoJoint with $T_{max} = 10$ outperforms all baselines on both datasets. And we have more detailed observations:

(1). Our multi-stage joint model is significantly superior to the baselines with single-stage unidirectional and bidirectional guidance in all metrics of both datasets. Compared with the unidirectional-guided state-of-the-art model SSRAN, InfoJoint achieves 2.7% improvement on Slot (F1), 2.0% improvement on Intent (Acc), 3.6% improvement on Overall (Acc) on the MixATIS dataset, and 0.3% improvement on Slot (F1), 1.1% improvement on Intent (Acc), 1.4% improvement on Overall (Acc) on the MixSNIPS dataset. Compared with the bidirectional-guided best model Co-guiding Net, InfoJoint also achieves significant and consistent performance improvements on all metrics.

(2). InfoJoint achieves a significant improvement in terms of overall accuracy. We could observe that the bidirectional models can perform better in overall accuracy than the unidirectional ones. This suggests that the bidirectional guidance achieves calibration and alignment of dual tasks, thereby obtaining better semantic parsing results. And InfoJoint ensures a gradual increase of beneficial information in bidirectional interaction through iterative enhancement. This gradually reduces cross-task propagation of erroneous semantics, thereby further facilitating sentence-level semantic analysis.

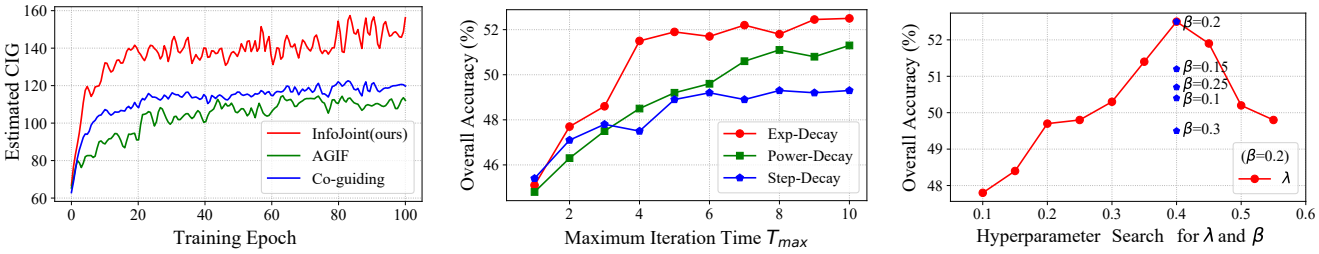
(3). We adopt a method similar to He and Garner (2023) to evaluate the performance of ChatGPT on these two datasets. As shown in Table 1, although ChatGPT has a strong ability for zero-shot learning in ID, it still lags far behind InfoJoint in overall accuracy. This difference suggests that ChatGPT may struggle to SF and comprehend the abstract connection between ID and SF. Hence, our work on joint multi-intent SLU remains of significant value to the community.

Ablation Study

We conduct a set of ablation experiments (shown in Table 2) to verify the effectiveness of our theoretical framework.

Effect of maximizing CIG. We conduct experiments to study the impact of MCCL on InfoJoint as illustrated in Table 2 with groups (a)(b)(c)(g). It can be seen that the performance of InfoJoint significantly decreases without MCCL to improve the cross-task information gain. Moreover, the lack of any level of contrastive loss ($\mathcal{L}_U$ or $\mathcal{L}_T$) can also affect prediction performance. This verifies that (1) MCCL is capable of extracting CIG sufficiently and effectively. (2) Maximizing CIG can effectively model semantic-level and word-level interactions of dual tasks to improve performance.

Effect of directional constraints. To further examine the effectiveness of margin penalty strategies, we show the ablation study on groups (d)(e)(g) in Table 2. We observe that the absence of any strategy ($\varphi_{mg}(t)$ or $\mathcal{L}_{mg}$), especially $\varphi_{mg}(t)$, can result in a significant decrease in predictive performance. This indicates that the margin penalty $\varphi_{mg}(t)$ and $\mathcal{L}_{mg}$ can effectively constrain the model to optimize in the positive direction ($i.e.$, satisfying Proposition 2).



(a) Dynamic analysis of information gain (b) Analysis of iteration and decay function (c) Hyperparamete tuning

Figure 4: Analysis Experiments of InfoJoint on MixATIS dataset (in color). (a) We conduct statistical analysis on the dynamic change of CIG during training on different models and find the universality of CIG optimization. (b) We analyze the effects of different decay functions and maximum iterations T_{max} on the performance of InfoJoint. (c) We demonstrate the impact of hyperparameters λ and β on model performance.

Effect of entropy weighting. By comparing the groups (f) and (g) of Table 2, we can further verify the effectiveness of the entropy weighting strategy in improving performance.

Method Analysis

Dynamic analysis of information gain. To further analyze the universality of the information-theoretic perspective and how InfoJoint works, We conduct an analysis experiment in MixATIS, which adopts a Monte Carlo estimator to obtain the CIG estimator of different models in each epoch (as shown in Fig. 4a). For AGIF and Co-guiding Net, we treat the output of the corresponding decoder as task random variables (*i.e.*, X_I and X_S), and discover that these two models implicitly optimize the CIG during training. And InfoJoint explicitly and significantly optimizes CIG to ensure deep alignment and interaction, resulting in better performance.

Analysis of iteration and decay function. We further analyze the effects of $\varphi_{mg}(t)$ and T_{max} as shown in Fig. 4b. We analyze three different types of functions: the exponential decay function ($a \cdot e^{-kt}$), the power function ($\frac{1}{1+t^{\gamma}}$), and the step function with an initial value of 0.5 and a decrease of 0.05 each time, where $a, k, \gamma = 0.9, 2, 3$. We find that the exponential function converges the fastest, possibly because it has the highest decay rate (gradient). We further observe that the larger T_{max}, the more sufficient cross-task interaction, and the higher the prediction performance.

Analysis of hyperparameter. As shown in Fig. 4c, we perform grid search on λ and β in the MixATIS dataset. We first fix $\beta = 0.2$ to balance MCCL and SLU tasks and observe that $\lambda = 0.4$ achieves the optimal balance between the main task and the information enhancement task. We then fix $\lambda = 0.4$ and find that $\beta = 0.2$ is the optimal trade-off.

Qualitative analysis. Following Zhu et al. (2023c), to better understand what the iterative interaction learns, we visualize the attention weight of InfoJoint and Co-guiding Net for comparison, as shown in Fig. 5. We observe that after sufficient iterations ($T_{max} = 10$), InfoJoint properly aggregates the intent `AddToPlaylist` at slots `add`, `song`, `to`, and `siesta`. This demonstrates InfoJoint successfully focuses weight on the correct slots during optimization and has a better interactive ability compared to prior methods.

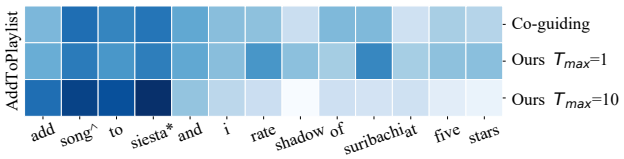


Figure 5: Attention heatmap in different approaches. $\wedge$ and $*$ denote specific slots `B-music` and `B-playlist`.

Experiments with Pre-training Model

Model	MixATIS	MixSNIPS
RoBERTa	49.7	80.2
AGIF+RoBERTa	50.0	80.7
SSRAN+RoBERTa	54.4	83.1
Co-guiding Net+RoBERTa	57.5	85.3
InfoJoint$_{T_{max}=10}$+RoBERTa	**58.6**	**86.1**
BERT	51.6	83.0
SSRAN+BERT	55.3	85.6
InfoJoint$_{T_{max}=10}$+BERT	**58.9**	**86.4**

Table 3: Overall (Acc) performance for pre-trained models with different architectures.

Following Qin et al. (2020); Cheng et al. (2023b), we use the pre-trained RoBERTa (Liu et al. 2019b) and BERT (Devlin et al. 2019) encoders to replace the original shared encoder. As shown in Table 3, InfoJoint outperforms all baselines when utilizing RoBERTa and BERT as encoders by a large margin, which further shows the universality and effectiveness of our iterative joint method on multi-intent SLU.

Conclusion

In this paper, we quantify the quality of joint interaction in multi-intent SLU from the perspective of information theory, and propose a principled framework with explainability and convergence. Based on this, we devise a novel joint model termed InfoJoint to model multi-stage dynamic interaction. Extensive experiments on two public datasets and analyses verify the effectiveness of InfoJoint.

Acknowledgments

We thank all anonymous reviewers for their constructive and insightful comments. This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No: GXWD20201231165807007-20200814115301001).

References

Bugliarello, E.; Mielke, S. J.; Anastasopoulos, A.; Cotterell, R.; and Okazaki, N. 2020. It's Easier to Translate out of English than into it: Measuring Neural Translation Difficulty by Cross-Mutual Information. In *Proc. of ACL*.

Cheng, L.; Yang, W.; and Jia, W. 2023. A scope sensitive and result attentive model for multi-intent spoken language understanding. In *Proc. of AAAI*.

Cheng, X.; Cao, B.; Ye, Q.; Zhu, Z.; Li, H.; and Zou, Y. 2023a. Ml-lmcl: Mutual learning and large-margin contrastive learning for improving asr robustness in spoken language understanding. In *Proc. of ACL Findings*.

Cheng, X.; Xu, W.; Zhu, Z.; Li, H.; and Zou, Y. 2023b. Towards spoken language understanding via multi-level multi-grained contrastive learning. In *Proc. of CIKM*.

Cheng, X.; Zhu, Z.; Li, H.; Li, Y.; Zhuang, X.; and Zou, Y. 2023c. Towards Multi-Intent Spoken Language Understanding via Hierarchical Attention and Optimal Transport. In *Proc. of AAAI*.

Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv preprint*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

E, H.; Niu, P.; Chen, Z.; and Song, M. 2019. A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling. In *Proc. of ACL*.

Gangadharaiah, R.; and Narayanaswamy, B. 2019. Joint Multiple Intent Detection and Slot Labeling for Goal-Oriented Dialog. In *Proc. of NAACL*.

Goo, C.-W.; Gao, G.; Hsu, Y.-K.; Huo, C.-L.; Chen, T.-C.; Hsu, K.-W.; and Chen, Y.-N. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proc. of NAACL*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. of CVPR*.

He, M.; and Garner, P. N. 2023. Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding. *ArXiv preprint*.

Hemphill, C. T.; Godfrey, J. J.; and Doddington, G. R. 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*.

Ji, B.; Zhang, T.; Zou, Y.; Hu, B.; and Shen, S. 2022a. Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective. In *Proc. of EMNLP*.

Ji, B.; Zhang, T.; Zou, Y.; Hu, B.; and Shen, S. 2022b. Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective. In *Proc. of EMNLP*.

Kim, B.; Ryu, S.; and Lee, G. G. 2017. Two-stage multi-intent detection for spoken language understanding. *Multimedia Tools and Applications*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of ICLR*.

Kurata, G.; Xiang, B.; Zhou, B.; and Yu, M. 2016. Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling. In *Proc. of EMNLP*.

Liu, Y.; Meng, F.; Zhang, J.; Zhou, J.; Chen, Y.; and Xu, J. 2019a. CM-Net: A Novel Collaborative Memory Network for Spoken Language Understanding. In *Proc. of EMNLP*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*.

Ni, J.; Young, T.; Pandelea, V.; Xue, F.; and Cambria, E. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *ArXiv preprint*.

OpenAI. 2023. ChatGPT.

Poole, B.; Ozair, S.; van den Oord, A.; Alemi, A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *Proc. of ICML*.

Qin, L.; Che, W.; Li, Y.; Wen, H.; and Liu, T. 2019. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proc. of EMNLP*.

Qin, L.; Liu, T.; Che, W.; Kang, B.; Zhao, S.; and Liu, T. 2021a. A co-interactive transformer for joint slot filling and intent detection. In *Proc. of ICASSP*.

Qin, L.; Wei, F.; Xie, T.; Xu, X.; Che, W.; and Liu, T. 2021b. GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling. In *Proc. of ACL*.

Qin, L.; Xu, X.; Che, W.; and Liu, T. 2020. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Proc. of EMNLP Findings*.

Song, M.; Yu, B.; Quangang, L.; Yubin, W.; Liu, T.; and Xu, H. 2022. Enhancing Joint Multiple Intent Detection and Slot Filling with Global Intent-Slot Co-occurrence. In *Proc. of EMNLP*.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proc. of NeurIPS*.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *Proc. of ICLR*.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous Graph Attention Network. In *Proc. of WWW*.

Wang, Y.; Shen, Y.; and Jin, H. 2018. A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In *Proc. of NAACL*.

Weld, H.; Huang, X.; Long, S.; Poon, J.; and Han, S. 2022. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding. *ACM Computing Surveys*.

Wu, D.; Ding, L.; Lu, F.; and Xie, J. 2020. SlotRefine: A Fast Non-Autoregressive Model for Joint Intent Detection and Slot Filling. In *Proc. of EMNLP*.

Xing, B.; and Tsang, I. 2022. Co-guiding Net: Achieving Mutual Guidances between Multiple Intent Detection and Slot Filling via Heterogeneous Semantics-Label Graphs. In *Proc. of EMNLP*.

Yao, K.; Peng, B.; Zhang, Y.; Yu, D.; Zweig, G.; and Shi, Y. 2014. Spoken language understanding using long short-term memory neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*.

Yeh, J. 2006. *Real analysis: theory of measure and integration second edition*. World Scientific Publishing Company.

Young, S.; Gašić, M.; Thomson, B.; and Williams, J. D. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*.

Zhang, C.; Li, Y.; Du, N.; Fan, W.; and Yu, P. 2019. Joint Slot Filling and Intent Detection via Capsule Neural Networks. In *Proc. of ACL*.

Zhang, X.; and Wang, H. 2016. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. In *Proc. of IJCAI*.

Zhu, Z.; Cheng, X.; Huang, Z.; Chen, D.; and Zou, Y. 2023a. Enhancing Code-Switching for Cross-lingual SLU: A Unified View of Semantic and Grammatical Coherence. In *Proc. of EMNLP*.

Zhu, Z.; Cheng, X.; Huang, Z.; Chen, D.; and Zou, Y. 2023b. Towards Unified Spoken Language Understanding Decoding via Label-aware Compact Linguistics Representations. In *Proc. of ACL Findings*.

Zhu, Z.; Xu, W.; Cheng, X.; Song, T.; and Zou, Y. 2023c. A dynamic graph interactive framework with label-semantic injection for spoken language understanding. In *Proc. of ICASSP*.