Uni-MIS: United Multiple Intent Spoken Language Understanding via Multi-View Intent-Slot Interaction

Shangjian Yin, Peijie Huang*, Yuhong Xu

College of Mathematics and Informatics, South China Agricultural University, China
sjy8460@163.com, {pjhuang, xuyuhong}@scau.edu.cn

Abstract

So far, multi-intent spoken language understanding (SLU) has become a research hotspot in the field of natural language processing (NLP) due to its ability to recognize and extract multiple intents expressed and annotate corresponding sequence slot tags within a single utterance. Previous research has primarily concentrated on the token-level intent-slot interaction to model joint intent detection and slot filling, which resulted in a failure to fully utilize anisotropic intent-guiding information during joint training. In this work, we present a novel architecture by modeling the multi-intent SLU as a multi-view intent-slot interaction. The architecture resolves the kernel bottleneck of unified multi-intent SLU by effectively modeling the intent-slot relations with utterance, chunk, and token-level interaction. We further develop a neural framework, namely Uni-MIS, in which the unified multi-intent SLU is modeled as a three-view intent-slot interaction fusion to better capture the interaction information after special encoding. A chunk-level intent detection decoder is used to sufficiently capture the multi-intent, and an adaptive intent-slot graph network is used to capture the fine-grained intent information to guide final slot filling. We perform extensive experiments on two widely used benchmark datasets for multi-intent SLU, where our model bets on all the current strong baselines, pushing the state-of-the-art performance of unified multi-intent SLU. Additionally, the ChatGPT benchmark that we have developed demonstrates that there is a considerable amount of potential research value in the field of multi-intent SLU.

Introduction

Spoken Language Understanding (SLU) plays a crucial role in task-oriented dialog systems, with the primary objective of constructing a semantic frame that encapsulates the user's request. This semantic frame is meticulously crafted through intent detection, identifying the user's intentions, and slot filling, extracting pertinent semantic elements. Since the two sub-tasks of intent detection and slot filling are closely tied (Tur and Mori 2011), dominant SLU systems adopt joint models to model the correlation between them (Liu and Lane 2016; Goo, Gao, and Hsu 2018; Qin, Che, and Li 2019).

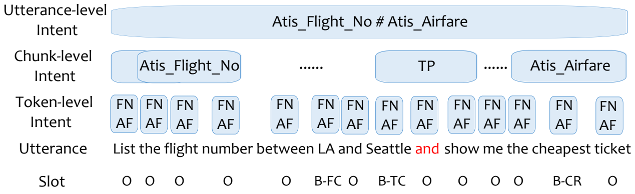

Utterance-level Intent
Atis_Flight_No # Atis_Airfare
Chunk-level Intent
Atis_Flight_No
......
TP
......
Atis_Airfare
Token-level Intent
FN AF
Utterance
List the flight number between LA and Seattle and show me the cheapest ticket
Slot
O O O O O B-FC O B-TC O O O O B-CR O

Figure 1: An example with multi-intent detection and slot filling: "FN", "AF", "FC", "TC", "CR" and "TP" denote "Flight_No", "Airfare", "froc.ctn", "toloc.ctn", "cost_relative" and "Transition Point". The multi-intent information can be categorised into three levels.

In real-life scenarios, users often express multiple intents within a single utterance, and the Amazon internal dataset showed that 52% of examples are multi-intent (Gangadharaiah and Narayanaswamy 2019). Figure 1 shows a two-intent example, which contains a classification task to classify the intent labels (i.e., predict the intents as : `Atis_Flight_No` and `Atis_Airfare`) and a sequence labeling task to predict the slot label sequence (i.e., label the utterance as {`O`, `O`, `O`, `O`, `O`, `B-frloc.ctn`, `O`, `B-toloc.ctn`, `O`, `O`, `O`, `O`, `B-toloc.ctn`, `O` }). However, most prior work only focused on the simple single-intent scenario, failing to effectively handle the multi-intent setting with the original network.

To deal with multi-intent scenarios, an increasing number of studies have begun to focus on modeling SLU in multi-intent settings. Xu and Sarikaya (2013) and Kim, Ryu, and Lee (2017) first explored the multi-intent SLU. Then Qin et al. (2020a) proposed an adaptive interaction framework (AGIF) to achieve fine-grained multi-intent information integration for token-level slot filling, which however suffers from information leakage issues due to its autoregressive architecture. Qin et al. (2021b) further proposed a global-locally graph interaction network (GL-GIN) to model slot dependency and interaction between multiple intents, which unfortunately potentially suffers from the fact that the intent information is misaligned as it simply treats intent detection as a token-level task. Recently, Huang et al. (2022) proposed a chunk-level intent detection (CLID) framework to split multi-intent into single-intent with an intent transi-

[*] Corresponding author.

tion point, achieving a promising performance. However, it is still possible to come across the problem of error propagation as it only utilizes the final predicted intent to guide slot filling and ignores rich anisotropic intent information during the join training state.

Most of the existing work has paid the major focus on how to accurately identify the intent and utilize the predicted token-level intent to guide the slot filling task. Few studies pay attention to building a fine-grained intent-slot interaction during the join training state. Indeed, anisotropic latent intent information plays a significant role in guiding slot filling. For example, as shown in Figure 1, utterance-level intent provides global semantic information and enables effective mitigation of information loss across the entire sentence, but its coarse-grained nature lacks the ability to capture detailed nuances; token-level intent offers a finer-grained interaction, allowing it to capture intent at the word level and achieve a more detailed understanding of the sentence's semantics and objectives. However, it may overlook successive fragments of intentional information in a sub-sentence; chunk-level intent considers the intent in segments, yet it encounters issues such as a lack of context and performance constraints. Considering these perspectives, the three views of intent-slot interaction complement each other, making it reasonable to combine them into a unified framework.

Based on the multi-view intent-slot interaction scheme, we further present a neural framework for unified multi-intent SLU (cf. Figure 2). First, Roberta (Liu et al. 2019b) is used to provide contextualized word representations and utterance intent information, and two BiLSTMs (Hochreiter and Schmidhuber 1997a) are used to generate the contextually sensitive hidden states for intent detection and slot filling, respectively. In the intent detection phrase, we adopt chunk-level intent detection (Huang et al. 2022) to get our final predicted intent. In the slot filling phrase, we inject multi-view intent-slot interaction to obtain a more comprehensive relationship between intent and slot. A fine-grain adaptive intent-slot graph interaction (Qin et al. 2020b) is finally used to get the slot filling result.

We conduct extensive experiments on two widely used benchmark datasets, MixATIS and MixSNIPS, and construct a ChatGPT evaluation benchmark for multi-intent SLU. The results show that our model outperforms current state-of-the-art (SoTA) methods while being able to effectively generalize from a single-domain dataset (MixATIS) to a multi-domain dataset (MixSNIPS), becoming the new SoTA method of unified multi-intent SLU.

In summary, the contributions of this work can be summarized as follows:

- We present an innovative method that casts unified multi-intent SLU as multi-view intent-slot interaction, where different levels of intent information are fully considered.
- We develop a neural framework for unified multi-intent, in which we propose a multi-view intent-slot information fusion for sufficiently capturing the different views of intent to guide slot filing.
- Our model pushes the current SoTA performance of multi-intent SLU on two widely used datasets on most evaluation metrics.
- We construct a ChatGPT evaluation benchmark for multi-intent SLU, showing that there is substantial room for improvement in performance, indicating significant potential for further advancements in this area.

Approach

In this section, we present a professionally crafted and logically coherent description of our proposed Unified Multi-Intent SLU Joint Learning Model (Uni-MIS), as illustrated in Figure 2. The Uni-MIS model comprises several key components: a shared Roberta encoder serving as the main encoder, two task-specific BiLSTM encoders, a chunk-intent detection decoder, a novel multi-view intent-slot interaction, and an adaptive intent-slot graph interaction. The central aim of our approach is to optimize intent detection and slot filling concurrently, employing a joint learning scheme for improved performance.

Shared Encoder

In our framework, Roberta (Liu et al. 2019b) is used as the encoder. Firstly, the user's utterances are tokenized using a tokenizer. Then, giving a sequence of words $T = \{x_1, x_2, ..., x_n\}$, the output representation is $E = \{e_1, e_2, ..., e_n\}$.

Task-Specific Encoder

The bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997b) have been successfully applied to sequence labeling tasks (Qin et al. 2021b). We adopt BiLSTM to read the input embedding $\{e_1, e_2, ..., e_n\}$ forwardly and backwardly to produce context-sensitive hidden states to promote its task-specific representation :

$$h_i^I = BiLSTM(e_i, h_{i-1}^I, h_{i+1}^I) \tag{1}$$

$$h_i^S = BiLSTM(e_i, h_{i-1}^S, h_{i+1}^S) \tag{2}$$

Chunk-Level Intent Detection

Users often express multiple intentions within a fragment within a sentence, not at the token-level or utterance-level. To this end, we adopt a chunk-level approach (Huang et al. 2022) to predict final intent. We employ a sliding window (SW) mechanism to capture contextual information within each chunk, facilitating regional intent detection. Using the SW, we anticipate the transition points in each utterance, allowing us to segment the utterance into sub-utterances, each with a singular intent. The intent of each sub-utterance is then determined by aggregating the intent predictions of the chunks (in a sliding window fashion) contained within it.

Sliding Window In the SW scheme, a window is used to slide through the utterance, and h_t^I is fed to calculate the $H_I^{win} = \{h_1^{win}, ..., h_w^{win}\}$, where w denotes the number of the window:

$$h_t^{win} = \sum_{i=1}^{win\_size} h_i^I \tag{3}$$

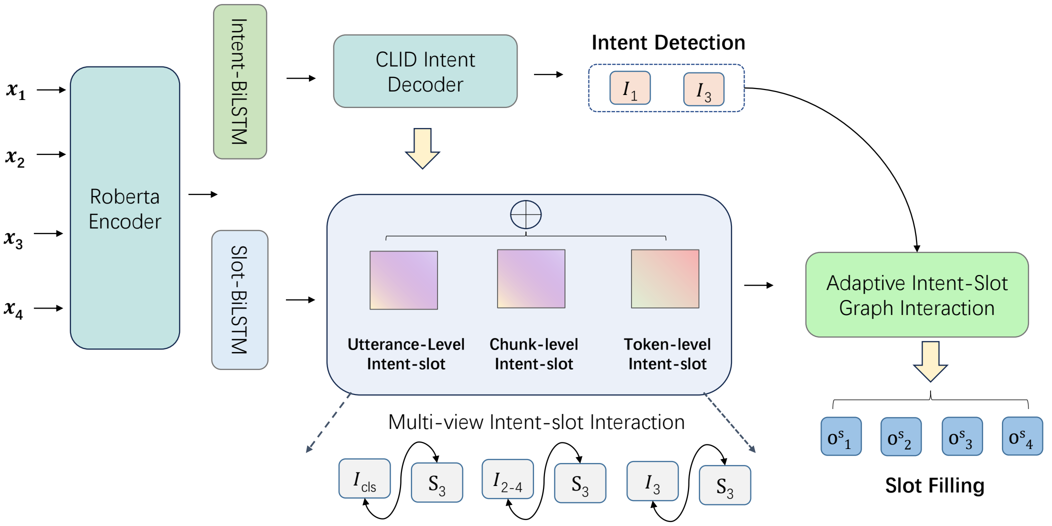


Figure 2: The overflow of model architecture and multi-view intent-slot interaction.

where $H_t^I = \{h_t^I, ..., h_{t+win\_size}^I\}$ is the matrix framed by a window to obtain the fragment intent information within a sentence.

Chunk-Intent Detection Derived from the sliding window, h_t^{win} is employed for detecting the intent of the chunk at the t-th window:

$$y_{I_t} = \sigma(W_I(\text{LeakyReLU}(W_h h_t^{win} + b_h)) + b_I) \tag{4}$$

$$o_t^I = \arg\max(y_t^I) \tag{5}$$

Where o_t^I represents the predicted intent label at the t-th window, σ denotes the sigmoid activation function, and W_h and W_I are trainable parameters. The terms b_h and b_I serve as bias parameters during training.

Multi-View Intent-Slot Interaction

The core contribution of this paper is the use of multi-view intent-slot interaction to relieve the problem caused by error propagation and utilize the multi-view intent information to guide slot filling. As shown in Figure 2, in the slot filling phrase, we inject utterance, chunk, and token-level intent interaction with slot-encoded status h_i^S.

View 1: Utterance Intent We treat [CLS] token produced by Roberta as the utterance intent:

$$I_{ut}^i = I_{cls} = MLP(Roberta(x_1, ..., x_n)) \tag{6}$$

View 2: Chunk Intent To align the slot token, the chunk level intent I_{ck}^i can be formulated:

$$I_{ck}^i = \begin{cases} I_{ck}^{i-1}, & \text{if } i > L - W + 1 \\ h_1^{win}, & \text{elif } i < W - 1 \\ h_{i-1}^{win}, & \text{else} \end{cases} \tag{7}$$

where L denotes the utterance length, and W denotes the window size as used in the Sliding Window (SW).

View 3: Token Intent The token-level intent is correspond to the representation of Intent-BiLSTM: $I_{tk}^i = h_i$.

Multi-View Intent-Slot Fusion Finally, we model the intent-slot relations in an utterance with utterance, chunk, and token-level intent-slot interaction and fuse them by the BiLSTM to capture the more diverse interaction between slot and intent, which can be formulated as:

$$h'^S_{I(v)} = BiLSTM(h_i^S \mid\mid I^i_{I(v)}) \tag{8}$$

$$h'^S_i = \sum^{view\_size} h'^S_{I(v)} \tag{9}$$

where $||$ denotes a concatenate operation, $h'^S_{I(v)}$ denotes the view of the hidden representation of intent-slot interaction, and h'^S_i denotes the final hidden representation of the slot resulting from the fusion of multiple views.

Adaptive Intent-Slot Graph Interaction

Instead of directly using s_t for predicting the slot label, we employ an adaptive intent-slot graph interaction (Qin et al. 2020a) to explicitly incorporate multi-intent information, guiding the prediction of the slot label at the t-th position. In this graph, the slot hidden state at time step t is denoted by s_t, and the predicted multiple intents information $I = \{I_1, \ldots, I_n\}$, where n denotes the number of predicted intents, are used as the initialized representations at time step t. The set $\tilde{H}^{[0,t]} = \{s_t, \phi_{\text{emb}}(I_1), \ldots, \phi_{\text{emb}}(I_n)\} \in \mathbb{R}^{(n+1)\times d}$ is constructed, where d denotes the dimension of the vertex representation, and $\phi_{\text{emb}}(\cdot)$ signifies the embedding matrix of intents. Moreover, the predicted intents are interconnected to account for their mutual interaction, as they all convey the intent of the same utterance.

With the L-layer adaptive intent-slot graph interaction, we obtain the final slot hidden state representation $\tilde{h}_0^{[L,t]}$ at time step t, which adaptive captures important intents information at the token-level. The representation $\tilde{h}_0^{[L,t]}$ is then utilized for slot filling:

$$y_t^S = \text{softmax}(W_s \tilde{h}_0^{[L,t]}) \tag{10}$$

$$o_t^S = \arg\max(y_t^S) \tag{11}$$

where o_t^S is the predicted slot label of the t-th word in the utterance.

Join Training

Taking into account the correlation between two sub-tasks, we opt for a joint training approach for our model. The objectives for chunk-level intent detection and slot filling are formulated as follows:

$$L_{\text{intent}} = -\sum_{i=1}^{n}\sum_{j=1}^{n_I} \hat{y}_i^{(j,I)} \log(y_i^{(j,I)}) \tag{12}$$

$$L_{\text{slot}} = -\sum_{i=1}^{n}\sum_{j=1}^{n_S} \hat{y}_i^{(j,S)} \log(y_i^{(j,S)}) \tag{13}$$

where n_I is the number of the intent, $\hat{y}_i^{(j,I)}$ is the gold intent label, n_S is the number of the slot and $\hat{y}_i^{(j,S)}$ is the gold slot label. The final joint objective is:

$$L = \alpha L_{\text{intent}} + L_{\text{slot}} \tag{14}$$

where α the weight parameter to balance the intent detection and slot filling tasks.

Experiments

Datasets

We conducted experiments on two publicly available multi-intent SLU datasets, namely MixATIS and MixSNIPS. The MixATIS dataset (Hemphill, Godfrey, and Doddington 1990; Qin et al. 2021b) is derived from the single-intent ATIS dataset, which is used to evaluate the performance of natural language understanding models. It consists of 13,162 training samples, 756 validation samples, and 828 testing samples, all originating from airline company queries. On the other hand, the MixSNIPS dataset (Coucke et al. 2018; Qin et al. 2021b) comprises queries from various domains such as restaurants, hotels, and movies. It is constructed from the single-intent SNIPS dataset and includes 39,776 training samples, 2,198 validation samples, and 2,199 testing samples. In both MixATIS and MixSNIPS datasets, the distribution of utterances with 1-3 intents is 30%, 50%, and 20%, respectively.

Experimental Settings

The batch size is 8 and 32 on MixATIS and MixSNIPS datasets, respectively. The dimensionality of the LSTM hidden units is 256. The number of multi-heads is 4 and 8 on the MixATIS and MixSNIPS dataset, respectively. The dimensionality of the intent-slot interaction hidden units is 256. The window size is 3. α is set to 0.1 The number of graph attention networks is set to 2. All layer numbers in the graph attention network are set to 2. The hyper-parameters are tuned using the validation set. We use Adam (Kingma and Ba 2015) to optimize the parameters in our model. For all the experiments, we select the model that works the best on the dev set and then evaluate it on the test set. All experiments are conducted on GeForce RTX 2080Ti and 3090Ti.

Baselines

We compare our model with the following baselines:

(1) Bi-Model (Wang, Shen, and Jin 2018): model the bi-directional relationship between intent detection and slot filling.

(2) Slot-Gated (Goo, Gao, and Hsu 2018): a slot-gated joint model to explicitly consider the correlation between slot filling and intent detection.

(3) SF-ID (E, Niu, and Chen 2019): establish a direct connection between the two tasks.

(4) Stack-Propagation (Qin, Che, and Li 2019): a stack-propagation framework to explicitly incorporate intent detection for guiding slot filling.

(5) AGIF (Qin et al. 2020b): an adaptive interaction network to achieve fine-grained multi-intent information integration.

(6) GL-GIN (Qin et al. 2021b): a local slot-aware and global intent-slot interaction graph framework to model the interaction between multiple intents and all slots within an utterance.

(7) SDJN (Chen, Zhou, and Zou 2022): a multiple instance learning and self-distillation framework for weakly supervised multiple intent information capturing.

(8) CLID (Huang et al. 2022): a chunk-level intent detection framework for recognizing intent within a fragment of an utterance.

(9) CLID(Roberta): a Roberta backbone model of CLID.

Main Results

We evaluate performance in these areas: F1 score is used for slot filling, accuracy for intent prediction, and overall accuracy for sentence-level semantic frame parsing according to Qin et al. (2021b) and Huang et al. (2022). The overall accuracy represents the proportion of sentences in which both intent and slot are accurately predicted and it is the most important metric in multi-intent SLU.

(1) As illustrated in Table 1, on the slot filling task, our framework outperforms the strong baseline in F1 scores on two datasets, which indicates the multi-view intent-slot interaction successfully utilizes the rich anisotropic intent information to guide the slot filling.

(2) Specifically, on MixATIS dataset, our framework outperforms the previous state-of-the-art model CLID (Roberta) by 3.1%, 2.4%, on sentence-level semantic frame parsing and slot filling respectively; on MixSNIPS dataset, it overpasses CLID (Roberta) by 1.2%, 0.2% and 0.4% on sentence-level semantic frame parsing, slot filling, and multiple intent detection, respectively. This is because our model utilizes the multi-view intent information to guide the slot filling, allowing the multiple intents to give a more comprehensive gaudiness and joint efficiency.

Model	MixATIS Dataset			MixSNIPS Dataset		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
Bi-Model (Wang, Shen, and Jin 2018)	83.9	70.3	34.4	90.7	95.6	63.4
Slot-Gated (Goo, Gao, and Hsu 2018)	87.7	63.9	35.5	87.9	94.6	55.4
SF-ID Network (E, Niu, and Chen 2019)	87.4	63.9	34.9	90.6	95.0	59.9
Stack-Propagation (Qin, Che, and Li 2019)	87.8	72.1	40.1	94.2	96.0	72.9
AGIF (Qin et al. 2020b)	86.9	72.2	39.2	93.8	95.1	72.7
GL-GIN (Qin et al. 2021b)	87.2	75.6	41.6	93.7	95.2	72.4
SDJN (Chen, Zhou, and Zou 2022)	88.2	77.1	44.6	94.4	96.5	75.7
CLID (Huang et al. 2022)	88.2	77.5	49.0	94.3	96.6	75.0
CLID (Roberta)	85.9	80.5	49.4	96.0	97.0	82.2
Uni-MIS (ours)	**88.3**	78.5	**52.5***	**96.4**	**97.2**	**83.4***

Table 1: SLU performance on MixATIS and MixSNIPS datasets. Values with * indicate that the improvement from our model is statistically significant over all baselines ($p < 0.05$ under t-test).

Model	intent num = 1			intent num = 2			intent num = 3		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
GL-GIN	88.0	91.3	72.6	87.3	76.2	39.1	86.8	63.1	23.0
CLID	88.6	94.7	76.4	88.1	77.5	48.4	87.6	64.3	28.5
CLID (Roberta)	88.6	95.8	77.6	85.4	80.3	48.8	84.7	66.8	29.0
Uni-MIS	89.2	95.1	**78.6**	87.6	78.3	**50.5**	86.7	66.7	**31.7**

Table 2: The result comes from the dataset MixATIS. The intent num denotes the number of intents in an utterance.

(3) Most importantly, our framework on most evaluation metrics achieves the state-of-the-art, showing a promising research direction for multiple intent spoken language understanding.

Analysis

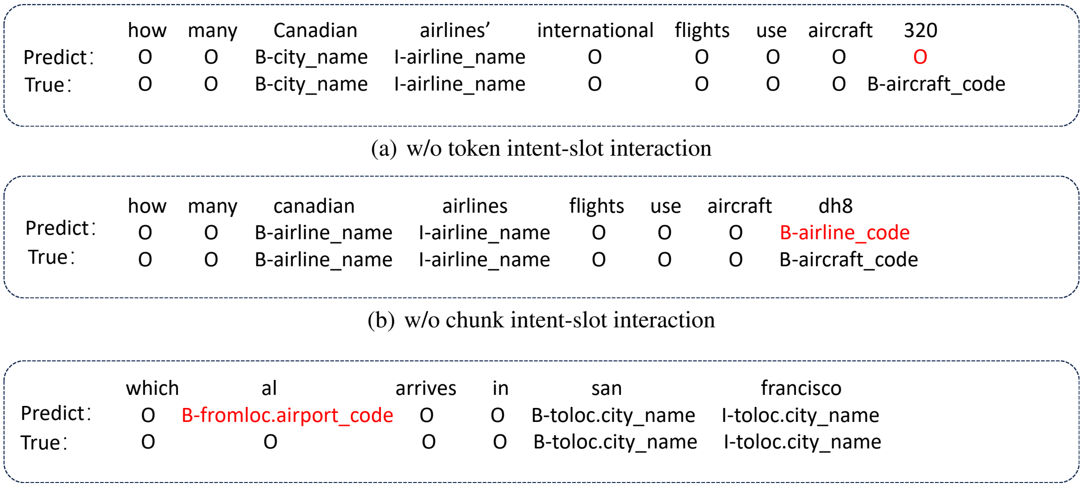
Improvement Analysis Our model outperforms the baseline models on most metrics for both the MixATIS and MixSNIPS datasets. To investigate the differences between the models, we have conducted a grouping based on the number of intents in the multi-intent datasets on MixATIS. From the Table 2, we can observe that as the number of intents increases, the model's scores gradually decrease. This indicates that both intent detection and slot filling become more challenging as the number of intents grows. However, our model consistently still outperforms the SoTA models across the most metrics, and its advantage becomes even more pronounced as the number of intents increases. With only one intent, our model's Overall(Acc) increases by 1.0%; And with two intents and three intents, it improves by 1.7% and 2.7% on Overall(Acc). This demonstrates that the multi-view intent interaction can effectively capture intent information and that modeling the joint process of intent detection and slot filling can enhance the overall accuracy. Specifically, we find out that the Slot(F1) decrease as the number of intents increases. We conclude that this is because multiple intent information could potentially introduce ambiguity in slot filling when predicted incorrectly.

Effectiveness of Utterance Intent-Slot Interaction To assess the impact of utterance intent-slot interaction, we conducted experiments where this interaction was removed from our multi-view intent-slot interaction model. The experimental results revealed a decrease of 1.4% and 1.3% in semantic parsing accuracy on two datasets, respectively, as shown in Table 3. These findings demonstrate that utterance-level intent-slot interaction plays a crucial role in providing global intent information within the utterance, thereby enhancing the correlational effect. The incorporation of global intent information effectively alleviates the problem of getting trapped in local optima during joint training.

Effectiveness of Chunk Intent-Slot Interaction To evaluate the effectiveness of chunk intent-slot interaction, we performed experiments by excluding it from our multi-view intent-slot interaction model. The results indicate a reduction of 0.6% and 1.3% in semantic parsing accuracy on two datasets, respectively, as shown in Table 3. This demonstrates that chunk-level intent-slot interaction contributes to offering fragmentary intent information within the utterance. It provides more localized intent-slot interaction information, thus enhancing the guidance of slot filling for fragment sentences.

Effectiveness of Token Intent-Slot Interaction To examine the effectiveness of token intent-slot interaction, we conducted experiments by eliminating it from our multi-view intent-slot interaction model. The experimental results show a decline of 1.0% and 1.3% in semantic parsing accuracy on two datasets, respectively, as shown in Table 3. This suggests that token-level intent-slot interaction provides specific intent information for individual tokens within the utterance, which strengthens the correlational effect. By offering more detailed intent-slot guidance, it enhances the flexibility of intent guidance for slots, thereby improving the overall join effect.

Case Analysis In our investigation, we amalgamate complementary multi-view intent-slot information and subsequently devise a network architecture that effectively harnesses diverse and directional intent cues to offer compre-


how many Canadian airlines' international flights use aircraft 320
Predict: O O B-city_name I-airline_name O O O O O
True: O O B-city_name I-airline_name O O O O B-aircraft_code
how many canadian airlines flights use aircraft dh8
Predict: O O B-airline_name I-airline_name O O O B-airline_code
True: O O B-airline_name I-airline_name O O O B-aircraft_code
which al arrives in san francisco
Predict: O B-fromloc.airport_code O O B-toloc.city_name I-toloc.city_name
True: O O O O B-toloc.city_name I-toloc.city_name

(a) w/o token intent-slot interaction

(b) w/o chunk intent-slot interaction

(c) w/o utterance intent-slot interaction

Figure 3: Case analysis.

Model	**MixATIS Dataset**			**MixSNIPS Dataset**		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
w/o utterance intent-slot	88.3	80.3	51.1	96.0	96.4	82.1
w/o chunk intent-slot	88.7	78.3	51.9	96.0	97.2	82.1
w/o token intent-slot	87.4	79.0	51.5	96.1	97.5	82.1
Uni-MIS (ours)	88.3	78.5	**52.5**	96.4	97.2	**83.4**

Table 3: Ablation experiments on the MixATIS and MixSNIPS datasets.

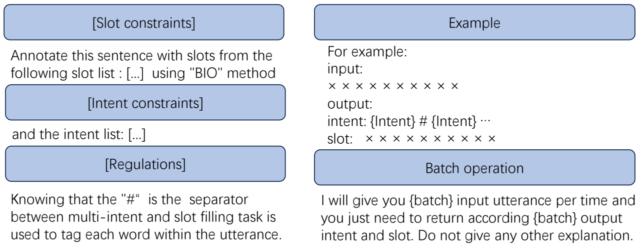
hensive guidance for slot filling tasks. The essence of this strategy is demonstrated in Figure 3 (a), where the absence of token intent-slot interaction results in the model's inability to correctly identify the specific number "320", erroneously categorizing it as an "O" tag in line with the preceding tag. A similar scenario is depicted in Figure 3 (b), where the omission of chunk intent-slot interaction leads the model to inaccurately label the term "dh8" as "B-aircraft_code" instead of the more appropriate "B-aircraft_code", thus failing to incorporate the contextual information from the preceding word "aircraft." Furthermore, Figure 3 (c) illustrates the implications of excluding utterance intent-slot interaction, causing the model to misclassify the term "al" as "B-fromloc.airport_code" rather than assigning it an "O" tag. This disregard for the holistic sentence context results in a loss of overarching information, consequently leading to semantic ambiguity. It is noteworthy that our proposed multi-view intent-slot interaction addresses these issues adeptly, yielding accurate outcomes in these cases. This substantiates the efficacy of our approach in jointly addressing both sub-tasks within the SLU framework.

Assessing Performance Using ChatGPT in Multi-Intent SLU Benchmark

In light of the swift advancements observed in the domain of large language models (LLMs), there arises a need to systematically assess their utility across conventional tasks. In pursuit of this goal, we construct an innovative benchmark centered around a multi-intent spoken language understanding task using the ChatGPT API ("gpt-3.5-turbo").

Figure 4 illustrates our devised prompt structure, which comprises five distinct components: slot constraints, intent constraints, regulations, examples, and batch operations. The slot and intent constraints serve the pivotal role of confining predictions within well-defined parameters. Meanwhile, the regulations facilitate enhanced comprehension of the specified format by ChatGPT. Accompanying these, we furnish a triad of example interactions (referred to as 3-shot examples), encompassing single-intent, dual-intent, and triple-intent scenarios. This strategic diversity aligns with the range of intents found in the dataset. Furthermore, we incorporate a batch operation mechanism to bolster assessment efficiency.

The resultant evaluation outcomes, as depicted in Figure 5, underscore the model's performance on the MixATIS and MixSNIPS datasets. Notably, the achieved scores are 0.36%, 5.91%, and 21.62% for sentence-level semantic frame parsing, slot filling, and multiple intent detection, respectively, in the context of MixATIS. Correspondingly, for MixSNIPS, the scores stand at 0.09%, 3.71%, and 71.49% for the same metrics. These results underscore a discernible gap in ChatGPT's proficiency when confronted with multi-intent SLU tasks. Intriguingly, ChatGPT exhibits superior performance in multiple intent detection within the MixSNIPS dataset as opposed to the MixATIS dataset. This divergence could be attributed to ChatGPT's adeptness in handling shallow,



[Slot constraints]

Annotate this sentence with slots from the following slot list : [...] using "BIO" method

[Intent constraints]

and the intent list: [...]

[Regulations]

Knowing that the "#" is the separator between multi-intent and slot filling task is used to tag each word within the utterance.

Example

For example:
input:
× × × × × × × × × ×
output:
intent: {Intent} # {Intent} ···
slot: × × × × × × × × × ×

Batch operation

I will give you {batch} input utterance per time and you just need to return according {batch} output intent and slot. Do not give any other explanation.

Figure 4: Example of prompt for multi-intent SLU.

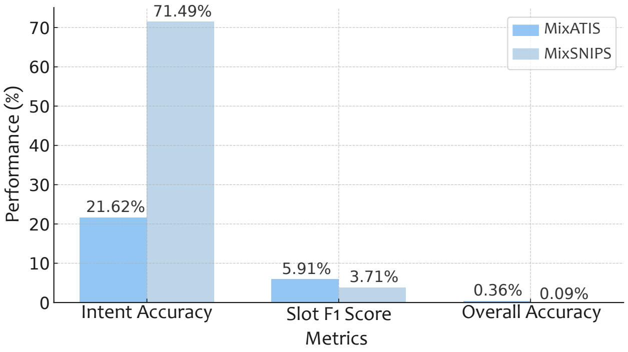


Figure 5: The ChatGPT performance within the two datasets.

multi-domain dialogues inherent to the MixSNIPS dataset. However, when faced with the intricate and singular-domain MixATIS, ChatGPT encounters formidable challenges.

Furthermore, with regard to slot filling, ChatGPT's performance is less favorable across both datasets. This decline in performance might be attributed to the model's relatively limited sequence labeling capabilities when confronted with the growing sequence lengths prevalent in multi-intent SLU tasks. Consequently, the accuracy of semantic frame parsing similarly dwindles to nearly negligible levels.

In light of these findings, it is imperative for us to pivot towards the development of intent-guided slot filling techniques and adopt joint training strategies within the context of multi-intent SLU. The incorporation of these innovations into the framework of LLMs and modern prompt technology is poised to yield significant enhancements in overall performance and efficacy.

Related Work

Intent Detection and Slot Filling

Intent detection and slot filling are often interrelated, giving rise to the development of integrated models that facilitate interaction between intent and slots. In recent years, techniques such as joint learning, which consider the strong correlation between intent and slots, have yielded outstanding results. Certain approaches to joint slot filling and intent detection involve the sharing of parameters (Liu and Lane 2016; Wang, Shen, and Jin 2018; Zhang and Wang 2016). The association between intent detection and slot filling can be modeled through unidirectional interaction or bidirectional-flow interaction (Qin et al. 2021c).

Unidirectional interaction approaches (Goo, Gao, and Hsu 2018; Li, Li, and Qi 2018; Qin, Che, and Li 2019) primarily focus on the flow from intent to slot. Gating mechanisms have been employed as specialized functions to guide slot filling (Goo, Gao, and Hsu 2018; Li, Li, and Qi 2018). Qin, Che, and Li (2019) proposed a token-level intent detection model to mitigate error propagation.

Bidirectional-flow interaction models (E, Niu, and Chen 2019; Zhang et al. 2019; Liu et al. 2019a; Qin et al. 2021a) consider the mutual impact between intent detection and slot filling. E, Niu, and Chen (2019) enhanced intent detection and slot filling bidirectionally through iteration mechanisms.

More recently, Chen, Zhou, and Zou (2022) introduced a Self-distillation Joint SLU model, leveraging multi-task learning. They also treated multiple intent detection as a weakly supervised challenge, employing Multiple Instance Learning (MIL). Cai et al. (2022) explicitly utilized the established association between slots and intents by connecting slots to their corresponding intents through a slot-intent classifier and intent-constrained attention. Huang et al. (2022) proposed a chunk-level intent detection framework, including an auxiliary task to identify intent transition points within utterances, thereby enhancing the recognition of multiple intents. The inferred intent information was then utilized to guide token-level slot filling.

Collectively, the aforementioned models comprehensively address intent detection, slot filling, and the incorporation of intent information for guiding slot filling at both the token and utterance levels.

Multi-View Learning

Multi-View learning (Zhao et al. 2017; Yan et al. 2021; Xu, Yu, and Chen 2022) has attracted substantial attention for its potential to enhance model performance by incorporating information from diverse data sources or feature perspectives. This approach recognizes that a comprehensive understanding of complex phenomena often arises from integrating different viewpoints. In multi-view learning, distinct sets of features or data representations are treated as separate "views" of the same underlying phenomenon. The key assumption is that each view offers a unique perspective that adds complementary information, leading to a more robust and accurate model. Various methods in multi-view learning include co-training, co-regularization, and consensus-based techniques. These strategies encourage the model to leverage the strengths of different views.

Conclusion

In this paper, we present an innovative method that casts unified multi-intent SLU as multi-view intent-slot interaction, where different levels of intent information are fully considered for sufficiently capturing the different views of intent to guide slot filing. Our model pushes the current SoTA performance of multi-intent SLU on two widely used datasets on most evaluation metrics. And we construct a ChatGPT evaluation benchmark for multi-intent SLU, showing that there is substantial room for further advancements in this area.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.71472068 and No.62306119) and the Natural Science Foundation of Guangdong Province (No.2021A1515011864).

References

Cai, F.; Zhou, W.; Mi, F.; and Faltings, B. 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*, 7607–7611. IEEE.

Chen, L.; Zhou, P.; and Zou, Y. 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*, 7612–7616. IEEE.

Coucke, A.; Saade, A.; Ball, A.; Bluche, T.; Caulier, A.; Leroy, D.; Doumouro, C.; Gisselbrecht, T.; Caltagirone, F.; Lavril, T.; Primet, M.; and Dureau, J. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

E, H.; Niu, P.; and Chen, Z. 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*, 5467–5471.

Gangadharaiah, R.; and Narayanaswamy, B. 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*, 564–569.

Goo, C.; Gao, G.; and Hsu, Y. 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)*, 753–757.

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, USA, June 24-27, 1990*. Morgan Kaufmann.

Hochreiter, S.; and Schmidhuber, J. 1997a. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hochreiter, S.; and Schmidhuber, J. 1997b. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Huang, H.; Huang, P.; Zhu, Z.; Li, J.; and Lin, P. 2022. CLID: A Chunk-Level Intent Detection Framework for Multiple Intent Spoken Language Understanding. *IEEE Signal Process. Lett.*, 29: 2123–2127.

Kim, B.; Ryu, S.; and Lee, G. G. 2017. Two-stage multi-intent detection for spoken language understanding. *Multim. Tools Appl.*, 76(9): 11377–11390.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Li, C.; Li, L.; and Qi, J. 2018. A Self-Attentive Model with Gate Mechanism for Spoken Language Understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 3824–3833.

Liu, B.; and Lane, I. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 685–689.

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 *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*, 1051–1060.

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. *CoRR*, abs/1907.11692.

Qin, L.; Che, W.; and Li, Y. 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*, 2078–2087.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 8193–8197.

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 *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*, 178–188.

Qin, L.; Xie, T.; Che, W.; and Liu, T. 2021c. A Survey on Spoken Language Understanding: Recent Advances and New Frontiers. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 4577–4584. ijcai.org.

Qin, L.; Xu, X.; Che, W.; and Liu, T. 2020a. AGIF: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1807–1816. Online: Association for Computational Linguistics.

Qin, L.; Xu, X.; Che, W.; and Liu, T. 2020b. Towards Fine-Grained Transfer: An Adaptive Graph-Interactive Framework for Joint Multiple Intent Detection and Slot Filling. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, 1807–1816.

Tur, G.; and Mori, R. D. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley, New York.

Wang, Y.; Shen, Y.; and Jin, H. 2018. A Bi-Model Based RNN Semantic Frame Parsing Model for Intent Detection and Slot Filling. In Walker, M. A.; Ji, H.; and Stent, A., eds., *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)*, 309–314. Association for Computational Linguistics.

Xu, P.; and Sarikaya, R. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, 78–83.

Xu, Y.; Yu, Z.; and Chen, C. L. P. 2022. Classifier Ensemble Based on Multiview Optimization for High-Dimensional Imbalanced Data Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Yan, X.; Hu, S.; Mao, Y.; Ye, Y.; and Yu, H. 2021. Deep multi-view learning methods: A review. *Neurocomputing*, 448: 106–129.

Zhang, C.; Li, Y.; Du, N.; Fan, W.; and Yu, P. S. 2019. Joint Slot Filling and Intent Detection via Capsule Neural Networks. 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*, 5259–5267.

Zhang, X.; and Wang, H. 2016. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. In Kambhampati, S., ed., *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2993–2999. IJCAI/AAAI Press.

Zhao, J.; Xie, X.; Xu, X.; and Sun, S. 2017. Multi-view learning overview: Recent progress and new challenges. *Inf. Fusion*, 38: 43–54.