Joint Intent Detection Model for Task-oriented Human-computer Dialogue System using Asynchronous Training

YIRUI WU, College of Computer and Information, Hohai University, China
HAO LI, College of Computer and Information, Hohai University, China
LILAI ZHANG, College of Computer and Information, Hohai University, China
DONG CHEN, College of Computer and Information, Hohai University, China
QIAN HUANG, College of Computer and Information, Hohai University, China
SHAOHUA WAN*, Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China

How to accurately understand low-resource languages is the core of the task-oriented human-computer dialogue system. Language understanding consists of two sub-tasks, i.e., intent detection and slot filling. Intent detection still faces challenges due to semantic ambiguity and implicit intentions with users' input. Moreover, separately modeling intent detection and slot filling significantly decrease the correctness and relevance between questions and answers. To address these issues, we propose a joint intent detection method using asynchronous training strategy. The proposed method firstly encodes local text information extracted by CNN and relationship information among words emphasized by attention structure. Later, a joint intent detection model with asynchronous training strategy is proposed by either fusing hidden states of intent detection and slot filling layers, or adopting the key information to fine-tune the whole network, greatly increasing the relevance of intent detection and slot filling subtasks. The accuracy achieved by the proposed method tested on an open-source airline travel dataset and a self-collected electricity service dataset, i.e., ATIS and ECSF, are 97.49% and 89.68% respectively, which proves the effectiveness of joint learning and asynchronous training.

CCS Concepts: • **Computing methodologies → Speech recognition**.

Additional Key Words and Phrases: Intent detection, Task-oriented human-computer dialogue system, Joint modeling, Asynchronous training

1 INTRODUCTION

In the past few decades, task-oriented human-computer dialogue system has attracted much attention due to its high availability and good market prospects. As shown in Fig. 1, a task-oriented human-computer dialogue system is usually composed of five modules: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management (DM), Natural Language Generation (NLG). SLU is the key to understanding user needs in task-oriented human-computer dialogue systems. Intent detection is the basis of spoken language comprehension

*Corresponding author.

Authors' addresses: Yirui Wu, wuyirui@hhu.edu.cn, College of Computer and Information, Hohai University, Nanjing, China; Hao Li, College of Computer and Information, Hohai University, Nanjing, China, lihao1998h@163.com; Lilai Zhang, College of Computer and Information, Hohai University, Nanjing, China, zhanglilai1999@gmail.com; Dong Chen, College of Computer and Information, Hohai University, Nanjing, China; Qian Huang, College of Computer and Information, Hohai University, Nanjing, China; Shaohua Wan, Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China, shaohua.wan@uestc.edu.cn.

2375-4699/2022/8-ART111 $15.00
https://doi.org/10.1145/3558096

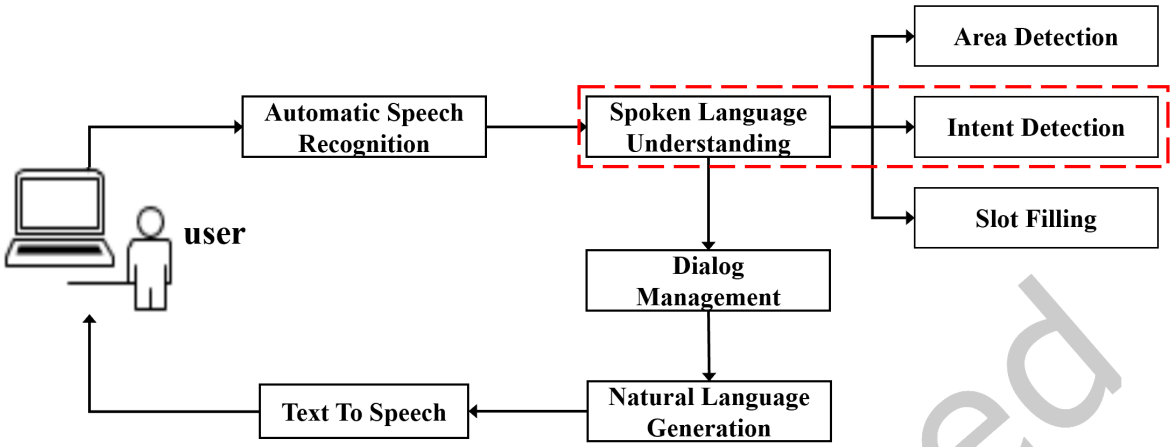

Fig. 1. Execution process of task-oriented human-computer dialogue system, Spoken Language Understanding is the key to understand user needs in task-oriented human-computer dialogue systems. Intent detection is the basis of spoken language comprehension tasks. It is responsible for understanding user needs and intentions, providing important supports for the dialogue management and dialogue generation module. The red dotted line marks the focus of our research.

tasks. It is mainly responsible for understanding user needs and intentions, and provides important support for the two modules of dialogue management and dialogue generation.

In practical application scenarios, the user's intention is often difficult to understand, for the following reasons: First, the user input information is more colloquial, using different expression methods to express the same intention; Second, the user input information is relatively short, some of the input information consists of only a few words; Third, user language expressions often carry pauses, mantras, and modal particles; Fourth, the semantics of Chinese is ever-changing, such as language ambiguity and implicit intentions. These problems make it difficult for the dialogue system to understand the user's real intention, resulting in ambiguity about the user's intention, and then leading to inaccurate detection results.

Existing deep learning intent detection algorithms can be mainly divided into CNN-based methods [17] and attention-based methods [23]. CNN-based methods can extract local feature information of text to achieve intent detection, but the location information between words is not well considered in the process of text feature extraction. While attention-based methods select important information features in the input sequence by introducing an attention mechanism. However, due to its own structure, it is impossible to obtain potential semantic information features, but these semantic information features are very important for intent detection. With the rapid increase in the volume of dialogue data from daily life, there is a growing demand for automatic dialogue system. Unfortunately, training a large dialogue system is generally infeasible due to the inadequacy of dialogue data with annotations, since creating large-scale dialogue datasets with annotations is costly and labor-intensive. This is the major challenge for low-resource computing.

Based on the above analysis, to combine the advantages of CNN and attention and capture the deep semantic information, we design a multi-dimensional feature fusion decoder on traditional Encoder-Decoder framework. This decoder is based on the attention module and CNN, and captures important semantic information from multiple angles. To avoid the loss of semantic information, the multi-dimensional representation of the input sequence is obtained by concatenating. Finally, intent detection is realized through the fully connected layer.

In recent years, to exploit the correlation between intent detection and slot filling tasks and avoid the problem of error propagation between sub-tasks in SLU, more and more researchers have proposed a joint modeling method to simultaneously improve the two sub-tasks. Liu et al. [25] proposed a joint model of Bi-RNN intent detection and slot filling based on Attention, which improved the accuracy of intent detection and the F1 value of slot filling to a certain extent, but there are still some problems with this model: First, the joint model simplifies the model structure by sharing the same coding layer, but the cross-influence between the hidden states of intent detection and slot filling tasks is not considered. Second, the influence of deep semantic information on input sequence is not considered in intent detection.

Therefore, we propose a joint modeling model of intent detection based on *asynchronous training*, which can capture more useful information and overcome the negative effects between the two tasks in a joint model. For the intent detection task, the attention module and convolutional neural network are introduced while fusing the slot filling hidden state, i.e., when considering the impact of the original input keyword information on the intent detection, the deep semantic information of the input sequence is obtained from multiple angles, The captured semantic information is concatenated to avoid loss of semantic information. At the same time, the semantic information is fused through the fully connected layer to realize the intent detection. In the slot filling task, the input alignment and attention mechanism are introduced, the relationship between the input sequence and the slot label is learned, and the slot label sequence of each position is output.

In fact, computational linguistics processing is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language. Our method is an example of computational linguistics processing, which is designed to perform the task of language computational understanding. Specifically, the propose method semantic understanding module analyze generated text instances of speech recognition module, thus expressing users intention. Since language is our most natural and most versatile means of communication, the proposed method would greatly facilitate our interaction with machines and software of all sorts as a typical method for computational linguistics processing.

Our adopted datasets, i.e., ATIS and ECSF, are special domain oriented dialogue datasets, i.e., Airline travel information and electricity custom service, where we prove the efficiency of the proposed method by conducting experiments on both low-resource datasets. Specifically, ATIS and ECSF datasets belong to English and Chinese languages. In self-collected ECSF datasets, we specially add some Cantonese languages, which is a variant version of mandarin but quite lack of related sample resources. Such experiments can be regarded as a trial to apply the proposed method on Indigenuous languages.

Specifically, we have handled the issue of "Indigenuous Languages" in three aspects. First, we construct our backbone network, i.e., BiGRU network, on a large dataset. After pre-training, we fine-tune our network on a small and indigenuous dataset, which offers capability of transforming knowledge of language understanding from large datase to small dataset, thus solving the issue of indigenuous language. Secondly, we build an attention scheme in network, which adjusts weights of neural network to be more task-specified for indigenuous languages. Thirdly, we have collected Indigenuous Language data to construct related dataset. In self-collected ECSF datasets, we own quantity of samples with Cantonese languages, where we handle it with sufficient information by training on such low-resource dataset.

The contribution of our work can be summarized as: 1) we propose a novel joint intent detection model based on CNN and attention structures, which involves local text information extracted by CNN and relationship information among words emphasized by attention structure to construct a smart task-oriented human-computer dialogue system. 2) we propose a joint intent detection model by asynchronously completing two tasks, i.e., either fusing hidden states of intent detection and slot filling layers, or adopting the key information to fine-tune the whole network, where the asynchronous training strategy greatly increases the relevance of intent detection and slot filling sub-tasks. 3) experiments conducted on an airline travel dataset and an electricity service dataset,

i.e., ATIS and ECSF, demonstrate that the proposed method has greatly improved the performance of intention detection tasks for special usage, compared with the existing methods. Ablation experiments show the benefits of our novelty-designed structures and the proposed asynchronous training strategy.

The rest of the paper is organized as follows. In Section 2, a brief overview of existing methods for intent detection models and slot filling models is introduced. In Section 3, the proposed multi-dimensional feature fusion intent detection model and joint model of intent detection based on asynchronous training are illustrated in detail. In Section 4, ablation experiments and comparative experiments are conducted on an airline travel dataset and an electricity service dataset. Finally, the paper will be summarized.

2 RELATED WORK

Spoken language understanding (SLU) not only plays an important role in task-oriented human-computer dialogue systems [9, 29, 39, 40], but also is a research focus of natural language processing (NLP). SLU mainly includes two sub-tasks: intent detection and slot filling [23, 33]. In specific domains, classification modeling and sequence tagging can be used for intent detection and slot filling tasks, respectively. There are two research methods, called independent modeling and joint modeling, that can be used for intent detection. The following paragraphs describe the recent research status of single modeling and joint modeling of intent detection, respectively.

2.1 Intent Detection Models

Intent detection is a crucial functionality of Natural Language Understanding (NLU) components in task-oriented human-computer dialogue systems [6, 29, 35]. As a matter of fact, the essence of intent detection is text classification [7, 45]. To understand the users current goal, the system must classify their utterance into several predefined intent classes.

In early studies on intent detection, the rule-based intent detection method requires the artificial construction of rule templates and category information to realize user intent detection. Its advantage is that the accuracy of intent detection can be high without a large amount of training data. However, Li et al. [22] found in their research that the size of the rule template would go up with the increase in the number of expression ways, requiring a large amount of manpower and resources to update the rule template, thus increasing the difficulty of manual maintenance of rules. Therefore, the rule-based approach can only meet the requirements of simple intent detection tasks.

With the significant improvement of computing ability and the rapid accumulation of data, natural language processing tasks is now mostly realized by deep learning. The intent detection model based on deep learning can effectively avoid the problems of rule-based and statistics-based intent detection. For example, some model based on LSTM has better performance than RNN [4, 5, 31, 36, 44]. In the intent detection experiment, Siwei et al. [28] found that encoding on Bi-GRU or Bi-LSTM has higher intent detection accuracy than on GRU or LSTM. More recently, the researchers mainly focus on the model's generalization performance under out-of-distribution (OOD) condition. For example, Ouyang et al. [30] propose using energy scores for unknown intent detection, which are theoretically well aligned with the density of the inputs, hence more suitable for OOD detection. Dopierre et al. [7] propose a meta-learning algorithm for short texts classification applied to the intent detection task. Offering the overview on existing methods for dialogue manager training with their advantages and limitations, Merdivan et al. [29] presents a new image-based method, which performs well and helps dialogue manager in expanding out of vocabulary dialogue tasks in comparison to Memory Networks. Shao et al. [33] present two novel frameworks, namely SA-CRFLV-I and SA-CRFLV-II, that use latent variables within random fields to make use of an encoding schema in the form of a latent variable, capturing the latent structure in the observed data. Since the distribution of outlier utterances is arbitrary and unknown in the training stage, Zhan et al. [45] propose a simple yet effective method to train an out-of-scope intent classifier in a fully end-to-end manner by simulating

the test scenario in training, which requires no assumption on data distribution and no additional post-processing or threshold setting.

On the other hand, current works on intent detection have been largely constrained only to English texts, and standard intent detection benchmarks also exist only in English texts [1, 20, 26]. The need to extend dialogue technology to other languages has only recently been recognized, so multilingual intent detection datasets are still very small: Schuster et al. [32] provide NLU data in three languages (English, Spanish, Thai), while a more recent MultiATIS++ dataset [43] manually translates the well-known ATIS dataset from English to 8 target languages. Gerz et al. [10] present a systematic study on multilingual and cross-lingual intent detection from spoken data, and release MINDS-14, a first training and evaluation resource for the task with spoken data, covering 14 intents extracted from a commercial system in the e-banking domain, with spoken examples available in 14 language varieties.

2.2 Joint Models of Intent Detection and Slot Filling

There are some problems in single intent detection modeling. First, it separates the relationship between the two tasks. Second, it will lead to the propagation, accumulation and amplification of errors. Therefore, more researchers [3, 38, 41] tend to build a joint model of intent detection task and slot filling task. The joint model simplifies the complexity of spoken language understanding to a certain extent, and it can also consider the correlation between the two tasks.

Much progress has been made in the research of slot filling. Yao et al. [44] used LSTM to solve it on the ATIS dataset, and introduced CRF scoring mechanism to improve the experimental results of sequence annotation. Kurata et al. [18] proposed to use encoder and decoder models for the task. Zhu et al. [47] introduced the attention mechanism into the encoder decoding model from sequence to sequence and achieved good experimental results.

Guo and Xu et al. [13] proposed a joint model experiment of intent detection and slot filling to avoid error propagation caused by pipeline method. Jeong m et al. [16] integrated the Maximum entropy model and the CRF [19] model and proposed a Triangular CRF (TriCRF) model for joint modeling of intent detection and slot filling, which achieved good experimental results in the mentioned tasks. Xu et al. [42] integrated deep learning and machine learning and proposed an experimental method combining CNN and TriCRF models. The model first uses the CNN model to extract dialogue text features and then transmits the feature information to the TriCRF model for intent detection and slot filling. The experimental results are better than TriCRF model, indicating that adding CNN model can improve the performance. Feng et al. [9] introduce the Evaluation of Chinese Human-Computer Dialogue Technology, which focuses on the identification of a user's intents and intelligent processing of intent words.

Recently, the Encoder-Decoder networks [34] have been successfully applied to many sequence learning problems, such as machine translation [27] and speech recognition [2]. Liu et al. [25] proposed a joint encoder-decoder network for intent detection and slot filling based on an attention module. The attention module here was introduced into the model to provide accurate focus and improve the accuracy of intent detection and the F1 value of slot filling. Goo et al. [11] proposed the Slot-gated model, which applied the intent information to the slot filling task and achieved excellent performance, but that did not explain how the slot filling task affected the intent detection task.

Focusing on the major challenge on how to use neural networks for extracting useful representations for each unit, Lin et al. [23] propose an attention segmental recurrent neural network (ASRNN) that relies on a hierarchical attention neural semi-Markov conditional random fields (semi-CRF) model for the task of sequence labeling. Zhang et al. [46] used CNN and GRU model to solve the two tasks of intent and slot filling. The model realizes the prediction of slot labels by learning the hidden state of GRU, and realizes intent detection by utilizing the maximum pooling in the CNN. Niu et al. [8] proposed an SF-ID network composed of an SF subnet and an ID

subnet. SF subnet applies intent information to slot filling task, while ID subnet applies intent information to slot filling task. The network model structure provides a two-way correlation mechanism for intent detection and slot filling. The experimental results show that the two-way correlation model can help the two subtasks promote each other.

Wang et al. [35] propose a novel Transformer encoder-based architecture with syntactical knowledge encoded for intent detection and slot filling. Specifically, they encode syntactic knowledge into the Transformer encoder by jointly training, which predicts syntactic parse ancestors and part-of-speech of each token via multi-task learning. Dopierre et al. [7] consider few-shot intent detection as a meta-learning problem, where the model is learning to learn from a consecutive set of small tasks named episodes. They thus propose ProtAugment, a meta-learning algorithm for short texts classification by extending Prototypical Networks, which limits overfitting on the bias introduced by the few-shots classification objective at each episode.

Compared with the above methods, our method is improved in the following two aspects: 1) In the intent detection task, a multi-dimensional feature fusion intent detection model based on Attention and CNN is proposed to effectively capture the deep semantic information input by users; 2) Meanwhile, to further study the influence of slot filling task on intent detection task, a joint model of intent detection based on asynchronous training is proposed based on the multi-dimensional feature fusion intent detection model.

3 METHOD

To effectively capture the deep semantic information input by users and the impact of slot filling task on intent detection task, a joint model for intent detection based on asynchronous training is proposed. In this section, we first introduce the overall architecture of the model, then introduce each of its components, and finally introduce our proposed asynchronous training strategy.

3.1 Overall Architecture

As shown in Fig. 2, the overall architecture of our joint model for intent detection is mainly composed of a temporal feature encoding layer, an intent detection decoding layer based on Attention and CNN, and a slot filling decoding layer based on aligned input and Attention.

The temporal feature encoding layer is composed of a word embedding layer and two bidirectional recurrent neural networks, which convert the input sequence into two hidden states corresponding to the intent detection task and slot filling task, respectively.

The intent detection decoding layer based on Attention and CNN is mainly composed of two parts: one is the contextual semantic information feature representation layer based on Attention; the other is the local semantic information feature representation layer based on CNN. To avoid the loss of semantic information features, the model will obtain the important semantic information feature of the context and the local semantic feature vector to be concatenated, and finally, realize the intent detection through the fully connected layer.

The slot filling task in spoken language understanding, i.e., sequence labeling, is usually implemented using a bidirectional recurrent neural network, but the relationship between the input sequence and the slot label cannot be captured during decoding, and the attention mechanism can dynamically pay attention to the relationship between input and output information during decoding.

3.2 Design of network structure

In this subsection, we introduce the temporal feature encoding layer, intent detection decoding layer, and asynchronous training strategy in detail.

Temporal feature encoding layer. After the words of the input sequence pass through the word vector embedding layer, they turn into $x_1, x_2, \ldots, x_T$ word vectors. The encoder of the bidirectional recurrent neural network is

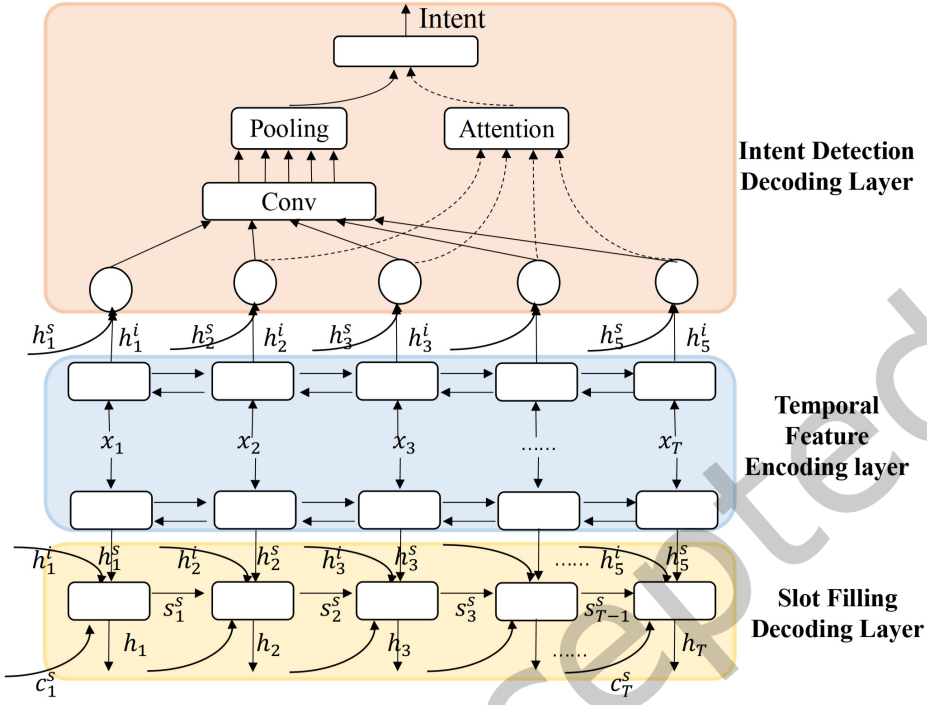

Fig. 2. Overall architecture of the joint model for intent detection

used to extract the hidden state at each moment in the input sequence. The recurrent neural network represents the extracted hidden state as $(\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_T})$, then the hidden state extracted by the bidirectional recurrent neural network at the current time t is $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$, which is about to get the forward and reverse hidden states are concatenated in series to obtain the final hidden state. Therefore, the hidden state captured by the intent detection task encoder is $(h_1^i, h_2^i, \ldots, h_T^i)$, and the hidden state captured by the slot filling task encoder is $(h_1^s, h_2^s, \ldots, h_T^s)$, where the superscripts i and s are used for intent detection task and slot filling task, respectively.

Intent detection decoding layer based on Attention and CNN is mainly composed of two parts: one is the contextual semantic information feature representation layer based on Attention; the other is the local semantic information feature representation layer based on CNN.

The decoding layer introduces an attention mechanism based on the time sequence feature encoding layer, calculates the weight of the time sequence feature, weights all the hidden states of the time sequence, and obtains the important semantic information feature representation of the context. The following formula is the process

of extracting contextually important semantic information features:

$$h_t = \left(h_t^i \oplus h_t^s\right) \tag{1}$$

$$e_t^i = W_h^i h_t \tag{2}$$

$$a_t^i = \frac{\exp\left(\left(s_t^i\right)^T e_t^i\right)}{\sum_{k=1}^{T} \exp\left(\left(s_t^i\right)^T e_k^i\right)} \tag{3}$$

$$d^i = \sum_{t=1}^{T} a_t^i h_t \tag{4}$$

$$D^i = d^i \oplus s_t^i \tag{5}$$

In the above equations, the superscripts i and s represent intent recognition and slot filling tasks, respectively; $\oplus$ represents concatenation; $h_t = (h_t^i \oplus h_t^s)$ is the result of concatenating hidden states in intent detection task and slot filling task; W_h^i represents the correlation measure weight in the intent detection task; a_t^i is the probability distribution of attention assigned to each input; s_t^i is the hidden state of the encoder at the last time; d^i represents the contextual semantic vector with attention mechanism in the intent detection task; D^i is the semantic information vector after the concatenation of the contextual semantic vector d^i with attention in the intent detection task and the hidden state s_t^i of the encoder at the last time.

The contextual semantic information feature representation layer based on attention can capture the important semantic information features of the context, but cannot obtain the potential semantic information features. To capture the local semantic information characteristics of the text with temporal special effects, a convolutional neural network is used based on recurrent neural network encoding.

In the local semantic information feature representation layer based on CNN, the influence of slot filling hidden states on the intent detection task should also be considered. The hidden states of the two tasks are concatenated $h_t = (h_t^i \oplus h_t^s)$ and output. Then, the sliding window of CNN with stride of 1 and size of m is used to extract the local features from $h_{1:M}^i, h_{2:m+1}^i, h_{3:m+2}^i, \ldots, h_{T-m+1:T}^i$ and get the feature map. Finally, a maximum pooling layer is used to sample the features extracted from the convolution layer to obtain the potential semantic information feature. It can be expressed as:

$$c_t^i = CNN(h_t : h_{t+m}) \tag{6}$$

$$C^i = (c_1^i, c_2^i, c_3^i, \ldots, c_{T-m+1}^i) \tag{7}$$

$$M^i = MaxPooling(C^i) \tag{8}$$

In the above equations, m is the size of sliding window. $c_1^i, c_2^i, c_3^i, \ldots, c_{T-m+1}^i$, means the calculated result of the sliding window of CNN on the hidden state $h_{1:m}, h_{2:m+1}, h_{3:m+2}, \ldots, h_{(T-m+1):T}$, and they make up the feature map C^i. Then M^i represents the local semantic information feature after a max pooling layer.

To avoid the loss of semantic information features, the model will obtain the semantic information feature D^i of the context and the local semantic feature vector M^i to be concatenated, and finally, realize the intent detection through the fully connected layer.

Slot filling decoding layer based on aligned input and attention mechanisms uses bidirectional recurrent neural network for decoding, for the decoder state at each moment, according to the decoder state s_{t-1}^s and predicted

label y_{t-1}^s at the previous moment, and the decoder hides at the current moment state $(h_t^s \oplus h_t^i)$ and context attention vector c_t^s, calculate the decoder state s_t^s at the current moment. It can be expressed as:

$$s_t^s = f\left(s_{t-1}^s, y_{t-1}^s, \left(h_t^s \oplus h_t^i\right), c_t^s\right), \tag{9}$$

where the superscripts i and s are used for intent detection task and slot filling task, respectively; $\oplus$ represents concatenation; $(h_t^s \oplus h_t^i)$ is the result of concatenating hidden states in slot filling task and intent detection task; c_t^s is the weighted sum of all hidden states, calculated by the formula shown below:

$$c_t^s = \sum_{t=1}^{T} a_t^s \left(\left(h_t^s \oplus h_t^i\right)\right) \tag{10}$$

$$a_t^s = \frac{\exp\left(e_t^s\right)}{\sum_{k=1}^{T} \exp\left(e_k^s\right)} \tag{11}$$

$$e_t^s = \left(s_{t-1}^s\right)^T W_h^s \left(h_t^s \oplus h_t^i\right) \tag{12}$$

In the above equations, a_t^s represents the attention distribution coefficient in the slot filling task; e_t^s represents the attention score in the slot filling task; W_h^i represents the correlation measure weight in the slot filling task; s_{t-1}^i represents the decoder state at the previous moment in the slot filling task.

For slot filling task, the slot decoder state is s_t^s at moment t, and the data is converted into the corresponding slot label category probability by Softmax function:

$$y_t^s = \text{Softmax}\left(W^s \cdot s_t^s + b^s\right). \tag{13}$$

3.3 Asynchronous Training Strategy

The proposed network uses the asynchronous training strategy to capture the cross-impact between the hiding states of intent detection and slot filling. To avoid its influence on the word embedding layer, the layer needs to be pretrained, and then the asynchronous training method is used to train the joint model. Fig. 3 shows our asynchronous training process, meanwhile Algorithm. 1 offers description on procedures of the whole asynchronous training strategy.

Asynchronous training algorithms are a popular way to reduce synchronization costs in large-scale optimization, and in particular for neural network training. However, for nonsmooth and nonconvex objectives, few convergence guarantees exist beyond cases where closed-form proximal operator solutions are available. As most popular contemporary deep neural networks lead to nonsmooth and nonconvex objectives, there is now a pressing need for such convergence guarantees. We present the propsoed asynchronous variants of training strategy and show that learning procedures have a stabilizing effect on training allowing to successfully train neural network.

1) Input a sequence into the intent detection model $M1$ and the slot filling model $M2$, and perform forward propagation to obtain the hidden state h_t^i and h_t^s respectively.

2) Freeze the encoding layer $f2$ of the slot filling model $M2$ and unfreeze the encoding layer $f1$ of the intent detection model $M1$. According to formulas (1) to (8), we can obtain semantic information feature D^i of the context and the local semantic feature vector M^i, and then concatenate them and output the result of intent detection through the fully connected layer. Then Optimize the intent detection model $M1$ through back propagation.

3) Similar to the previous step, we then freeze the encoding layer $f1$ of the intent detection model $M1$ and unfreeze the encoding layer $f2$ of the slot filling model $M2$. According to formulas (9) to (13), we can output the result of slot filling. Then Optimize the slot filling model $M2$ through back propagation.

4) Repeat 2) and 3) the loss function converges.

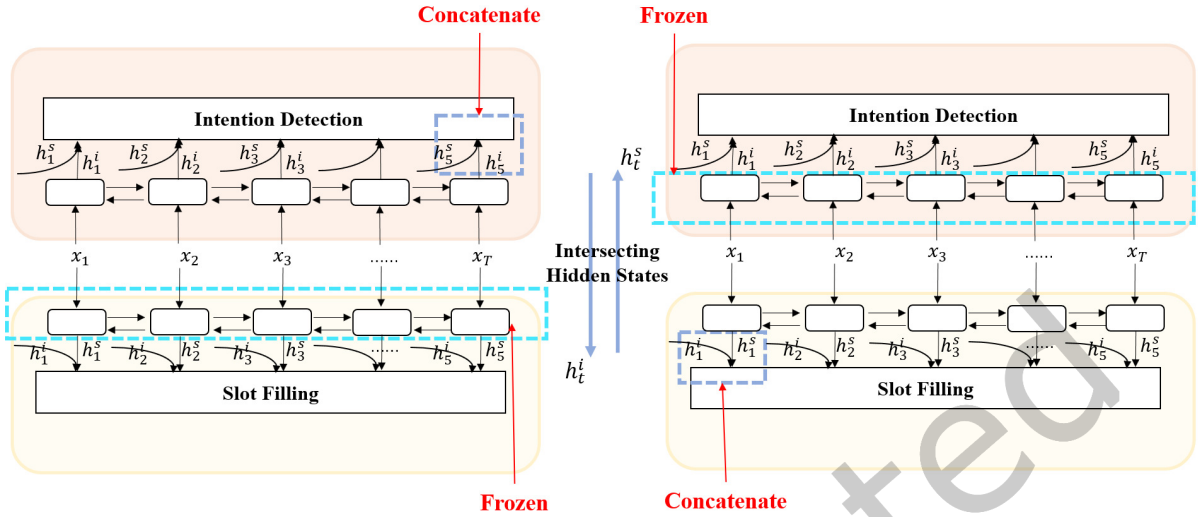

Fig. 3. Asynchronous Training Strategy, where the left part shows asynchronous training process of intention detection training in the proposed joint model, and the right part is the process of slot filling training.

Algorithm 1: Asynchronous Training Strategy

input : batch data B
output: optimized intent detection model M_1,
optimized slot filling model M_2

while *loss function does not converge* **do**
  select a small batch of data B;
  $h_t^i \leftarrow M_1(B)$ // getting the hidden state;
  $h_t^s \leftarrow M_2(B)$ // getting the hidden state;
  freeze the encoding layer f_2 *of* M_2;
  unfreeze the encoding layer f_1 *of* M_1;
  $p_1 \leftarrow f_1 \oplus h_t^s$ // concatenating;
  pass p_1 *to the decoder* g_1 *of* M_1;
  optimize M_1 *through back propagation*;
  freeze the encoding layer f_1 *of* M_1;
  unfreeze the encoding layer f_2 *of* M_2;
  $p_2 \leftarrow f_2 \oplus h_t^i$ // concatenating;
  pass p_2 *to the decoder* g_2 *of* M_2;
  optimize M_2 *through back propagation*;

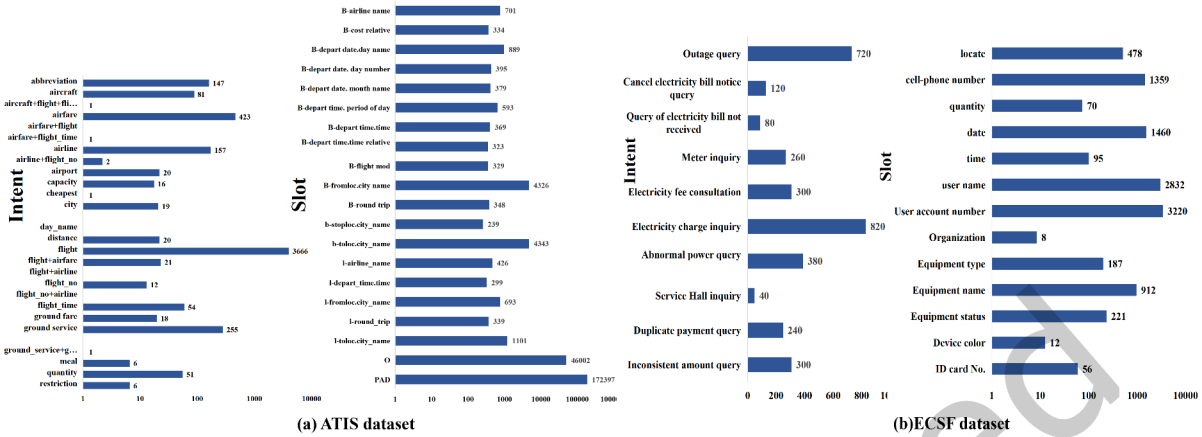

Fig. 4. Distribution of intents and slots in ATIS(a) and ECSF(b) dataset

ATIS									
Sentence	Show	flights	from	dallas	to	baltimore	in	first	class
Intention	flight								
Slot	O	O	O	B-fromloc.city_name	O	B-toloc.city_name	O	B-class_type	I-class-type
Mandarin Chinese sample in ECSF									
Sentence	我(I)	想 (Want to)	查询 (Inquiry)	12	月份 (December)	的	电费 (Electricity Charge)	使用 (Usage)	情况 (Situation)
Intention	电费查询(Electricity Charge Inquiry)								
Slot	O	O	O	B-date_month	O	O	O	O	O
Cantonese sample in ECSF									
Sentence	请(Please)	问(Tell me)	最近(the last)	10	天(Day)	嘅	用电 (Electricity)	乜样(Usage)	
Intention	电使用查询(Electricity Inquiry)								
Slot	O	O	O	B-date_month	O	O	O	O	

Fig. 5. Examples of labels for questions in ATIS and ECSF datasets. It's noted that the question in ECSF dataset are asked in either Mandarin Chinese or Cantonese, where Cantonese can be regarded as Indigenuous and low-resource language with insufficient sample resources.

4 EXPERIMENTS

In this section, we show the effectiveness of the proposed model for intent detection task. We first introduce two datasets. Then, we describe experimental details. Finally, the effectiveness of our proposed method is proved through ablation experiments and comparative experiments.

Table 1. Results of ablation experiments with different combination of feature extraction methods on ATIS and ECSF

Method	Accuracy(%)	
	ATIS	ECSF
CNN	94.98	83.80
LSTM	94.05	83.10
BiLSTM	94.49	86.59
GRU	94.27	84.08
BiGRU	95.40	86.91
BiLSTM-Attention	95.94	88.30
BiLSTM-CNN	95.85	87.24
BiLSTM-Attention-CNN	96.11	88.78
BiGRU-Attention	95.89	87.45
BiGRU-CNN	96.40	89.01
BiGRU-Attention-CNN	**96.62**	**89.36**

4.1 Datasets

We use the Airline Travel Information System Dataset (ATIS) and the Electricity Customer Service Field Dialogue Dataset (ECSF) in this experiment. The ATIS dataset is the open-source English data on Kaggle. The ECSF data set is collected from real phone recording data collected by the customer service center and converted into Chinese text data. Fig. 4 shows the distribution of intents and slots in ATIS and ECSF.

Specifically, ATIS (Airline Travel Information System) is an open-source English dataset often used in researches on oral comprehension and intent detection. The ATIS data set adopts the BIO labeling method: B-xxx indicates that the word is the first position of a certain slot value, I-xxx indicates that the word is the middle or the end of a certain slot value, and O indicates that the word is not a slot value. In the experiment, we normalize the ATIS dataset, add PAD for filling and UNK tags for unregistered words to the vocabulary, and add the slot tag corresponding to the word PAD (also denoted by PAD) to the slot tag, UNK It is indicated by the O label and does not belong to any slot type label. There are 945 words in the ATIS dataset, 4978 samples in the training set, and 893 samples in the validation set, which contain 26 intents and 130 slot labels.

ECSF (Electricity Customer Service Field Dialogue Dataset) collects Chinese question-and-answer text data between manual customer service and users in the electricity bill field of the customer service center. At the same time, the laboratory members where the intern company is located jointly complete the data labeling work, and screen out the data that meets the intent identification and slot filling task experiments. The ECSF data set in this article has a total of 3260 samples, which contains 10 intents and 13 slot tags. It's noted that Fig. 5 show examples of intents and slot labels for questions in ATIS and ECSF datasets, where one slot corresponds to a word.

4.2 Ablation experiments

We conduct two ablation experiments to prove the effectiveness of multi-dimensional feature fusion decoder and asynchronous training strategy for joint model respectively. Table 1 shows the comparison results with different combinations of BiRNN(BiLSTM and BiGRU), attention module and CNN. Table 2 show the comparison results of single intention detection model and asynchronous trained joint model.

As shown in Fig. 6 (a), the BiLSTM-Attention model and BiGRU-Attention model have higher intent detection accuracy than BiLSTM and BiGRU models by 1.45% and 0.71%, respectively, indicating that attention mechanism in intent detection can better obtain important information features in the input sequence, which is helpful to

Table 2. Results of ablation experiments with single intention detection model and asynchronous trained joint model on ATIS and ECSF

Method	Accuracy(%)		F1 Score(%)	
	ATIS	ECSF	ATIS	ECSF
BiLSTM-Attention-CNN	96.11	88.78	NA	NA
BiGRU-Attention-CNN	96.62	89.36	NA	NA
Our model (BiLSTM)	97.18	89.45	96.02	88.71
Our model (BiGRU)	**97.49**	**89.68**	**96.49**	**89.38**

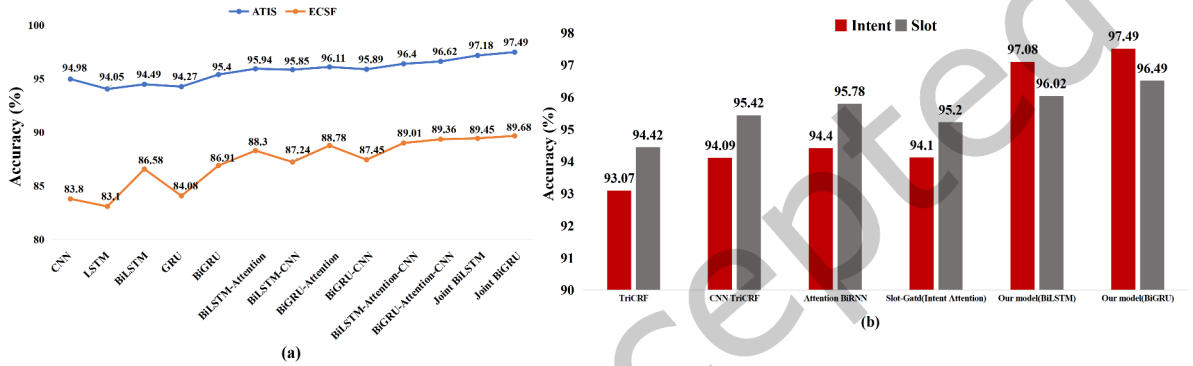

Fig. 6. Plot of comparisons in performance, where (a) refers to the accuracy achieved by the proposed method with different ablation methods, and (b) compares performance in both intent detection and slot filling tasks achieved by different joint models on ATIS dataset.

identify user's intent. At the same time, both bilSTM-CNN model and BiGRU-CNN model have higher accuracy of intent detection than the BiLSTM model and BiGRU model, indicating that more complete semantic information features can be extracted by considering the temporal and local features of input sequences in the process of intent detection.

Among them, BiGRU with attention and CNN has the highest accuracy of intent detection on ATIS dataset and ECSF dataset, 96.62% and 89.36% respectively, indicating that the model proposed in this chapter can capture important semantic information of the input sequence from multiple angles, and fusion of the obtained multi-dimensional features can effectively improve the accuracy of intent detection.

In addition, intent detection joint model with BiLSTM encoder (Our model, BiLSTM) is 0.78% more accurate than the single model with BiLSTM encoder (BiLSTM-Attention-CNN). Intent detection joint model with BiGRU encoder (Our model, BiGRU) is 0.87% higher than the accuracy of the single model with BiGRU encoder (BiGRU-Attention-CNN). So it is verified that the hidden state of slot filling task is introduced in the intent detection task, and a higher intent detection accuracy can be obtained.

4.3 Comparison experiments

We conduct some comparative experiments on other joint models and the proposed asynchronous trained joint model on ATIS dataset. Table 3 shows the accuracy of intent detection and the F1 value of slot filling of each model.

Table 3. Comparison performance among the proposed method and other joint models on ATIS dataset

Model	Accuracy (%)	F1 Score (%)
TriCRF [16]	93.07	94.42
CNN TriCRF [42]	94.09	95.42
Attention BiRNN [25]	94.40	95.78
Slot-Gated(Intent Attention) [11]	94.10	95.20
Joint Seq [15]	92.6	94.3
Attention BiRNN [24]	91.1	94.2
Slot-Gated Intent [12]	93.6	94.8
Self-Attentive [21]	96.8	95.1
Bi-Model[37]	96.4	95.5
SF-ID Network [14]	96.6	95.6
Our model(BiLSTM)	97.18	96.02
Our model(BiGRU)	**97.49**	**96.49**

It can be seen from Fig. 6 (b) that the intent detection accuracy and slot filling F1 value of Our model (BiLSTM) are 97.18% and 96.02%, respectively, while the intent detection accuracy and slot filling F1 value of Our model (BiGRU) are 97.49% and 96.49%, respectively. The proposed joint model of intent detection based on asynchronous training has higher accuracy and F1 value than other joint models. At the same time, in the ATIS dataset, the intent detection joint model based on asynchronous training with BiGRU encoder has higher accuracy than the model with BiLSTM encoder.

Comparing with comparative studies, we can find the proposed method significantly improve performance on all three measurements. Being good at dealing with both English and Chinese dataset, the proposed method still achieves better performance than other methods on ATIS dataset. In fact, most existing NLP methods prefer to independently model two subtasks or only focus to well performing only one subtask, which leads to the ignorance of iterative information. By introducing and well describing interactive information, the proposed dual structure has proved its effectiveness by experimental results.

To observe the convergence speed of the proposed model, as shown in Fig. 7, we show the variation of the intent detection accuracy with the number of iterations on the ATIS dataset. It can be found that the proposed model has converged in about 3 rounds and behaves much smoother, which shows that the proposed model has a better convergence effect compared with other models. This fast convergence and stability are very important properties for the task-oriented human-computer dialogue system.

4.4 Experimental details

In this section, we will introduce the setup on the experiments. For ATIS dataset, the maximum input length accepted by the model is 50; the embedding size of words is 250; the hidden layer vector size of the decoding layer is set to 128; the batch size is set to 128; and the number of training epochs is 30. For ECSF dataset, the maximum input length accepted by the model is 100; the embedding size of words is 300; the hidden layer vector size of the decoding layer is set to 128; the batch size is set to 64; and the number of training epochs is 40. Adam optimization algorithm is used to update parameters in the network. The initial learning rate r is set to 0.005, and the exponential decay rates β_1, β_2 are set to 0.9, 0.999 respectively. Our experiment is based on the Keras framework on Intel Iris Plus Graphics 645 graphics card.

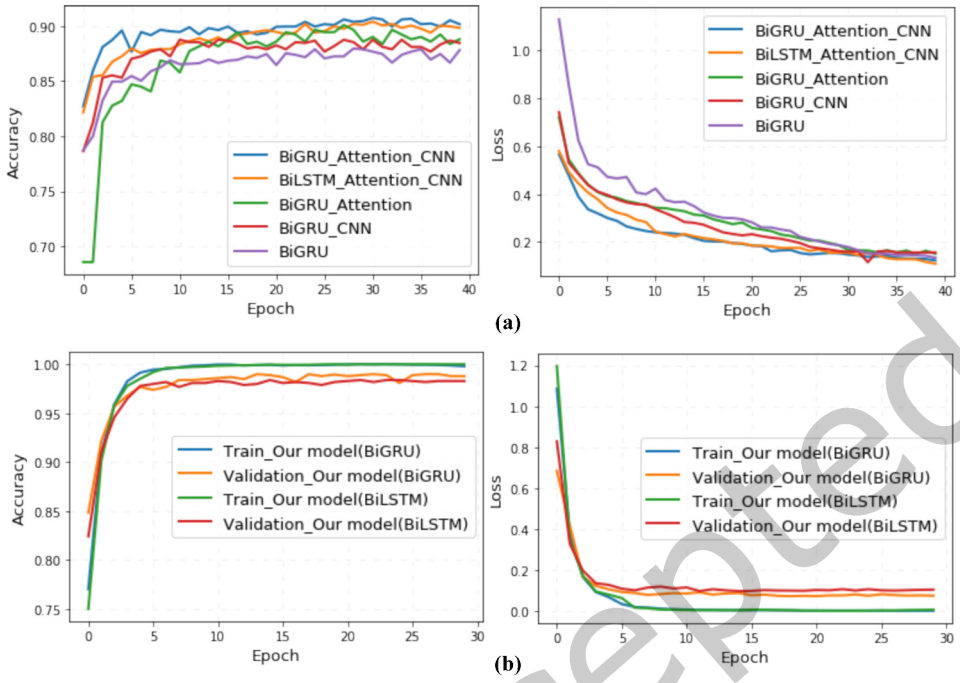

Fig. 7. The plot of accuracy and loss achieved by different models on ATIS dataset

5 CONCLUSION

This paper first researches single modeling of intent detection in spoken language understanding. To effectively capture the deep semantic information input by users, a multi-dimensional feature fusion decoder based on attention and CNN is proposed. Then this paper researches joint modeling of intent detection and slot filling. A joint model of intent detection based on asynchronous training is proposed. This model introduces the hidden state of the slot filling task in the intent detection task. The proposed joint model is tested on the Chinese and English datasets and shows that it can effectively improve the accuracy of intent detection. In total, the model proposed in this paper performs well on the ATIS dataset and ECSF dataset, verifying the effectiveness of the model.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 62172438.

REFERENCES

[1] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient Intent Detection with Dual Sentence Encoders. *CoRR* abs/2003.04807.

[2] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 4960–4964.

[3] Chen Chen, Jiange Jiang, Yang Zhou, Ning Lv, Xiaoxu Liang, and Shaohua Wan. 2022. An edge intelligence empowered flooding process prediction using Internet of things in smart city. *J. Parallel and Distrib. Comput.* 165 (2022), 66–78.

[4] Chen Chen, Lei Liu, Shaohua Wan, Xiaozhe Hui, and Qingqi Pei. 2021. Data dissemination for industry 4.0 applications in internet of vehicles based on short-term traffic prediction. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–18.

[5] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of Empirical Methods in Natural Language Processing*. 551–561.
[6] Alice Coucke, Alaa Saade, et al. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces. *CoRR* abs/1805.10190.
[7] Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. ProtAugment: Intent Detection Meta-Learning through Unsupervised Diverse Paraphrasing. In *Proceedings of Association for Computational Linguistics*. 2454–2466.
[8] 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 Association for Computational Linguistics*. 5467–5471.
[9] Zixian Feng, Caihai Zhu, et al. 2021. An Evaluation of Chinese Human-Computer Dialogue Technology. *Data Intell.* 3, 2, 274–286.
[10] Daniela Gerz, Pei-Hao Su, et al. 2021. Multilingual and Cross-Lingual Intent Detection from Spoken Data. In *Empirical Methods in Natural Language Processing*. 7468–7475.
[11] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 753–757.
[12] 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, Volume 2 (Short Papers)*. 753–757.
[13] Zhaohan Daniel Guo, Gökhan Tür, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *Proceedings of IEEE Spoken Language Technology Workshop*. 554–559.
[14] E Haihong, 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 Annual Meeting of the Association for Computational Linguistics*. 5467–5471.
[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 *Proceedings of Annual Conference of the International Speech Communication Association*. 715–719.
[16] Minwoo Jeong and Gary Geunbae Lee. 2008. Triangular-Chain Conditional Random Fields. *IEEE Trans. Speech Audio Process.* 16, 7, 1287–1302.
[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of Empirical Methods in Natural Language Processing*. 1746–1751.
[18] Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging Sentence-level Information with Encoder LSTM for Semantic Slot Filling. In *Proceedings of Empirical Methods in Natural Language Processing*. 2077–2083.
[19] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of International Conference on Machine Learning*. 282–289.
[20] Stefan Larson, Anish Mahendran, et al. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Empirical Methods in Natural Language Processing*. 1311–1316.
[21] Changliang Li, Liang Li, and Ji Qi. 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*. 3824–3833.
[22] Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Nat. Lang. Eng.* 12, 3, 229–249.
[23] Jerry Chun-Wei Lin, Yinan Shao, Youcef Djenouri, and Unil Yun. 2021. ASRNN: A recurrent neural network with an attention model for sequence labeling. *Knowl. Based Syst.*, 106548.
[24] Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454* (2016).
[25] Bing Liu and Ian R. Lane. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Proceedings of International Speech Communication Association*. 685–689.
[26] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking Natural Language Understanding Services for Building Conversational Agents. In *Proceedings of International Workshop on Spoken Dialog System Technology*. 165–183.
[27] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of Empirical Methods in Natural Language Processing*. 1412–1421.
[28] Shengfei Lyu and Jiaqi Liu. 2021. Convolutional Recurrent Neural Networks for Text Classification. *J. Database Manag.* 32, 4, 65–82.
[29] Erinc Merdivan, Deepika Singh, Sten Hanke, and Andreas Holzinger. 2018. Dialogue Systems for Intelligent Human Computer Interactions. In *1st Workshop on Behavioral Change and Ambient Intelligence for Sustainability and 2nd Workshop on Affective Interaction with Avatars and Robots*. 57–71.
[30] Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Energy-based Unknown Intent Detection with Data Manipulation. In *Proceedings of Association for Computational Linguistics*. 2852–2861.
[31] Suman V. Ravuri and Andreas Stolcke. 2015. Recurrent neural network and LSTM models for lexical utterance classification. In *Proceedings of International Speech Communication Association*. 135–139.

[32] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In *Proceedings of Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3795–3805.
[33] Yinan Shao, Jerry Chun-Wei Lin, Gautam Srivastava, Alireza Jolfaei, Dongdong Guo, and Yi Hu. 2021. Self-attention-based conditional random fields latent variables model for sequence labeling. *Pattern Recognit. Lett.*, 157–164.
[34] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of Neural Information Processing Systems*. 3104–3112.
[35] 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 *Proceedings of AAAI Conference on Artificial Intelligence*. 13943–13951.
[36] Yu Wang. 2017. A new concept using LSTM Neural Networks for dynamic system identification. In *Proceedings of America Control Coference*. 5324–5329.
[37] Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235* (2018).
[38] Yirui Wu, Haifeng Guo, Chinmay Chakraborty, Mohammad Khosravi, Stefano Berretti, and Shaohua Wan. 2022. Edge computing driven low-light image dynamic enhancement for object detection. *IEEE Transactions on Network Science and Engineering* (2022).
[39] Yirui Wu, Wenxiang Liu, and Shaohua Wan. 2021. Multiple attention encoded cascade R-CNN for scene text detection. *Journal of Visual Communication and Image Representation* 80 (2021), 103261.
[40] Yirui Wu, Yuntao Ma, and Shaohua Wan. 2021. Multi-scale relation reasoning for multi-modal Visual Question Answering. *Signal Processing: Image Communication* 96 (2021), 116319.
[41] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot User Intent Detection via Capsule Neural Networks. In *Proceedings of Empirical Methods in Natural Language Processing*. 3090–3099.
[42] Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*. 78–83.
[43] Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-End Slot Alignment and Recognition for Cross-Lingual NLU. In *Empirical Methods in Natural Language Processing*. 5052–5063.
[44] Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *Proceedings of IEEE Spoken Language Technology*. 189–194.
[45] Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y. S. Lam. 2021. Out-of-Scope Intent Detection with Self-Supervision and Discriminative Training. In *Proceedings of Association for Computational Linguistics*. 3521–3532.
[46] Xiaodong Zhang and Houfeng Wang. 2016. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding. In *Proceedings of International Joint Conference on Artificial Intelligence*. 2993–2999.
[47] Su Zhu and Kai Yu. 2017. Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 5675–5679.