Neurocomputing 413 (2020) 348–359

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

End-to-end masked graph-based CRF for joint slot filling and intent detection

Hao Tang, Donghong Ji [*], Qiji Zhou

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China

ARTICLE INFO

Article history:
Received 25 December 2019
Revised 13 April 2020
Accepted 29 June 2020
Available online 9 July 2020
Communicated by Zhi Yong Liu

Keywords:
Slot filling
Graph-based CRF
Intent detection
Mask mechanism
End-to-end structure

ABSTRACT

Slot filling and intent detection are the basic and crucial fields of natural language processing (NLP) for understanding and analyzing human language, owing to their wide applications in real-world scenarios. Most existing methods of slot filling and intent detection tasks utilize linear chain conditional random field (CRF) for only optimizing slot filling, no matter the method is a pipeline or a joint model. In order to describe and exploit the implicit connections which indicate the appearance compatibility of different tag pairs, we introduce a graph-based CRF for a joint optimization of tag distribution of the slots and the intents. Instead of applying the complex inference algorithm of traditional graph-based CRF, we use an end-to-end method to implement the inference, which is formulated as a specialized multi-layer graph convolutional network (GCN). Furthermore, mask mechanism is introduced to our model for addressing multi-task problems with different tag-sets. Experimental results show the superiority of our model compared with other alternative methods. Our code is available at https://github.com/tomsonsgs/e2e-mask-graph-crf.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Joint multi-task of slot filling and intent detection is a popular research field of natural language understanding (NLU) aimed at identifying the semantic slots and the corresponding intent, which can be considered as a summary of the entire sentence. NLU is critical to the overall performance of goal-oriented spoken dialogue systems such as GoogleHome, AmazonEcho, and TmallGenie. As shown in Fig. 1, we demonstrate a typical sample from the training set of Snips, in which slot filling task is labeled by BIO tag-set and intent detection is an utterance-level classification task.

Recently recurrent neural network (RNN) [1] based methods, especially gated recurrent unit (GRU) [2] and long short-term memory (LSTM) [3], have already shown great ability when processing sequence labeling and sentence classification problems. For the joint optimization of NLU, several joint models [4–6] are proposed to enhance the connection between the intent and the slots located in the same utterance. These models exploit the different joint approaches of NLU, and the results demonstrate that joint model obtains better performance compared with pipeline methods. Attention mechanism [7] improves the performance of a large number of various natural language processing (NLP) tasks. Thus, many attention-based methods [4,5] are proposed to model the connection between words and intents or slots. These methods are specially designed for NLU considering the influence of different slots. Furthermore, Zhang et al. [8] propose a joint Capsule Neural Network (CapsNN) to enhance the description of inner connection between word-slot and slot-intent pairs. Recently, a basic Bidirectional LSTM (BiLSTM) network [9] equipped with Bidirectional Encoder Representations from Transformers (BERT), has achieved a great improvement and obtained the best performance among all released methods on Atis and Snips datasets [10].

Linear chain condition random field (LC–CRF) [11] is widely used in numerous sequence labeling tasks, including the slot filling task. LC–CRF is also applied to the joint NLU task, for the purpose of only optimizing the performance of slot filling. However, LC–CRF is not effective for the NLU task. There are two major shortcomings of LC–CRF: I. The output tag distribution of each cell is influenced by only surrounding cells within the distance usually set as one or two; II. LC–CRF can only be implemented on slot filling (sequence labeling) owing to the restriction of chain structure. In NLU task, slots and intents are implicitly connected rather than independent, the appearance of slot-slot pairs and slot-intent pairs are associated and can affect each other. For example, the intent ‘PlayMusic’ is highly connected with the slot ‘album’ but it conflicts with the slot ‘airplane_name’, and the slot ‘album’ is also against the appearance of the slot ‘airplane_name’.

[*] Corresponding author.
E-mail addresses: tanghaopro@whu.edu.cn (H. Tang), dhji@whu.edu.cn (D. Ji), qiji.zhou@whu.edu.cn (Q. Zhou).

https://doi.org/10.1016/j.neucom.2020.06.113
0925-2312/© 2020 Elsevier B.V. All rights reserved.

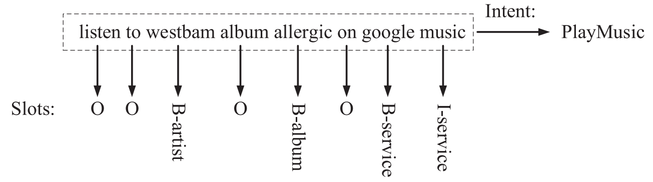


Fig. 1. Overall demonstration of a typical sample of joint slot filling and intent detection task, where BIO tags are used for slot filling.

To address this problem, we introduce a graph-based CRF to handle the NLU tasks. For NLU tasks with only slot filling, we use a word-level fully-connected graph to construct a graph-based CRF module, which indicates that all word-level slot tags are connected and associated with each other. For joint NLU, we develop two forms of graph-based CRF, i.e. semi-connected and fully-connected graphs. Thus, the intent tags and slot tags are interactive in our joint model. Graph-based CRF is not commonly used in NLP field owing to the complexity of the learning and inference of Probability Graph Model (PGM). We introduce an end-to-end (E2E) graph-based CRF to simplify the learning and inference process. For joint NLU task, we introduce the mask mechanism to solve the incompatibility between slot tags and intent tags. Our model is based on the BERT enhanced BiLSTM, which also contains the attention module for intent detection.

We list our four main contributions as follows:

- We introduce a graph-based CRF to our model for further modeling the implicit connection in slot-slot and slot-intent pairs.
- We design a graph-based CRF for not only NLU tasks with only slot filling, but also joint NLU tasks with slots and intents.
- We use an end-to-end graph-based CRF to accelerate the learning and inference process, and utilize the mask mechanism to cope with joint tasks. Our model achieves the best performance on all four datasets compared with other state-of-the-art (SOTA) alternatives.
- We develop an E2E graph-based CRF which is suitable and versatile for all word-level or utterance-level classification tasks or multi-tasks. The module is convenient to implement and extremely suitable for multi-tasks. We think that this is a great contribution to our NLP community, by upgrading the traditional LC–CRF to an E2E graph-based CRF, which is more general and powerful.

2. Related work

Intent detection is a sub-field of text classification, specialized for intent purpose. With the recent development of deep learning, many neural network based models [12–15] are proposed to identify the specific intent from various diversely expressed utterances of natural language. Recently, various capsule-based text classification models [16–18] are proposed that aggregate word-level features for utterance-level classification via dynamic routing-by-agreement. Joint capsule neural network proposed and modified by Zhang et al. [8] is applied for joint NLU and achieves great performance.

Currently, slot filling is usually treated as a sequential labeling task. Recurrent neural networks such as Gated Recurrent Unit (GRU) or Long Short-term Memory Network (LSTM) are used to learn context-aware word representations, and Conditional Random Fields (CRF) [19] is used to annotate each word based on its surrounding slot types. Recently, Attentional methods [20,21] introduce the self-attention mechanism for CRF-free sequential labeling.

For joint learning of slot filling and intent detection, joint models are proposed to solve two tasks simultaneously in a unified framework. Xu et al. [22] propose a Convolution Neural Network (CNN) based sequential labeling model for joint NLU. Hakkani et al. [23] adopt a Recurrent Neural Network (RNN) network to slot filling, and the last hidden state of the RNN is used to predict the utterance intent. RNN-based encoder-decoder models [4,5] further improve the performance of joint slot filling and intent detection. An attention weighted sum of all encoded hidden states is used to predict the utterance intent. Goo et al. [24] utilize slot-gated mechanism as a special gate function in LSTM to improve slot filling by the learned intent context vector. Furthermore, Zhang et al. [8] propose a joint Capsule Neural Network which uses a dynamic routing-by-agreement schema between capsule layers. It explicitly models the hierarchical relationship between words and slots, as well as intents on the utterance-level via dynamic routing-by-agreement. Note that most of the models mentioned above are already equipped with LC–CRF, which actually provides minimal contribution to the performance improvement.

On the other hand, many modern pre-trained methods such as Glove [25] and Word2Vec [26], which are specially trained on a large corpus for unlabeled language generation, have shown an impressive performance improvement in SQL parsing tasks [27]. Recently developed contextualized word representations such as ELMO [28] and BERT [29] show superior performances in many natural language processing (NLP) tasks. The outputs from final Transformer [30] block are used in all these models as the hidden representation of the clauses. BERT uses PieceTokenizer to slice each word into pieces, we further develop a sum-based aggregation operation to restore the pieces back to their father tokens. BERT enhanced BiLSTM [10] is also used for joint NLU, which achieves the best performance among all SOTA alternatives.

3. Preliminaries

In this part, we first simply introduce the basic definition and the computation procedure of graph-based CRF, then we give a brief illumination of NLU tasks.

3.1. Graph-based CRF

As shown in Fig. 2, assume that there are N nodes in the graph. n_i denotes the i-th node and $u_i \in R^k$ denotes the unary tag potential of i-th node, where k is the tag size. We denote binary tag potential between the i-th and j-th nodes as $b_{i,j}$. Following equations can be formulated when a training sample $\{X, Y\}$ is given, where $X = \{n_0, \ldots\}$ and $Y = \{t_0, \ldots\}$, and t_i is the tag of i-th node:

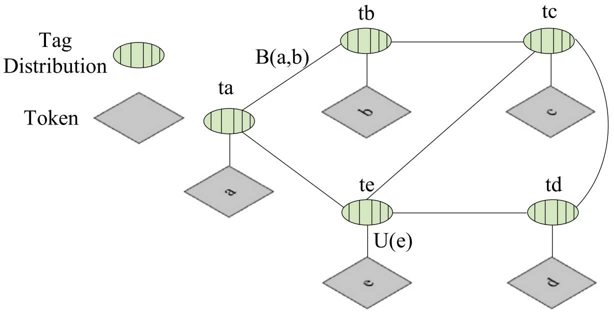


Fig. 2. A detailed demonstration of a simple graph-based CRF with five unique nodes, which are not fully connected.

$$u_i = U(X_i, w_U), \tag{1}$$
$$b_{i,j} = B(X_i, X_j, w_B), \tag{2}$$
$$logit_{Y,X} = exp\left(\sum_i \sum_j^{j\in adj(i)} w_i^b b_{i,j}(Y_i, Y_j) + \sum_i w_i^u u_i(Y_i)\right), \tag{3}$$
$$Z_X = \sum_Y logit_{Y,X}, \tag{4}$$
$$P(Y|X) = logit_{Y,X}/Z_X, \tag{5}$$

where $adj(i)$ is the adjacent nodes of i-th node. U is the unary potential function and B is the binary potential function. w_U and w_B are the inner parameters of U and B. Thus, the learning target is $argmax_{w^u,w^b,w_U,w_B}(P(Y|X))$ when $\{X, Y\}$ is given, w^u, w^b, w_U, w_B are the parameters that require optimization. During the inference process, after all parameters are determined, we need to choose the best Y in all possible tag combinations to maximum $P(Y|X)$ when only X is given in testing set. Precise inference of graph-based CRF is a NP-hard problem, loopy belief propagation (LBP) algorithm [31] is proposed to perform the efficient inference of graph with cycles. Note that LBP is only an approximate inference algorithm and can be considered as an extension of belief propagation (BP).

3.2. Problem definition

Given a natural language sentence $X = \{w0, w1, \ldots\}$, our model outputs the BIO tags of each word to form the corresponding slots, and also the intent tag of the entire sentence. Our model is then evaluated under both slot filling and intent detection performance.

4. End-to-end masked graph-based CRF (E2EMG-CRF)

The overall structure of our model is demonstrated in Figs. 3 and 4, we enhance the encoder with BERT and propose a joint model for NLU task. We use BiLSTM with attention as our basic model. Four forms of CRF are performed in our experiment. Except for applying traditional LC–CRF for optimizing the slot filling, we mainly test another three proposed graph-based CRF models including graph-based CRF for only slot filling, semi-connected and fully-connected graph-based CRF for joint NLU tasks.

4.1. BERT Enhanced Utterance Encoder

Bidirectional Encoder Representations from Transformers (BERT) has already shown great ability on word-level representations for multiple tasks of different domain. For each word in the sentence, we use PieceTokenizer to further slice the word into pieces. As shown in Fig. 5, sliced words will later be restored to their original integrated words by simply applying sum operation over piece representations. We conduct our experiment with the base version of BERT.

We use w_k to represent the k-th word. Assume that all tokens are processed by BERT, then hd_k denotes the k-th word ERT embedding. Bidirectional LSTMs [9,3] are applied for the word embeddings after BERT:

$$\widehat{hd}, \overline{hd} = BiLSTM\left(\left[\overrightarrow{hd}, \overleftarrow{hd}\right]\right), \tag{6}$$

where $hd = \{hd_0, \ldots\}$ is the inputs, and → and ← on the top represent the forward and backward order of original embeddings, respectively. $\widehat{hd} = \left\{\widehat{hd}_0, \ldots\right\}$ is the outputs of $BiLSTM$. $\overline{hd}$ is the last state of BiLSTM encoder.

4.2. Self attention

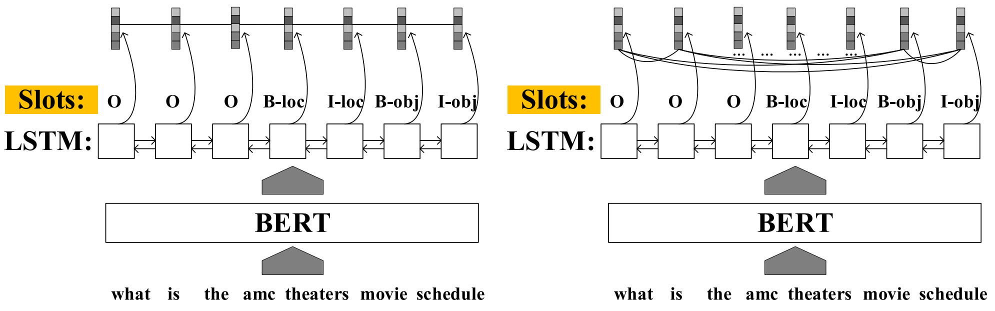
Intent is often triggered by specific words in the sentence. Considering that different words have different influence on intent detection, self attention mechanism is used in our model to dynamically assign associated weight to seqfuential words. The weight reflects the word-level contribution to formulate the final intent. The complete computation process is listed below:

$$c_k = MLP_2\left(\left[\widehat{hd}_k, \overline{hd}\right]\right), \tag{7}$$
$$\hat{a} = softmax(c), \tag{8}$$
$$h^{att} = \sum_{k=0}^{N-1} \hat{a}_k \widehat{hd}_k, \tag{9}$$
$$h = MLP_2\left(\left[h^{att}, \overline{hd}\right]\right), \tag{10}$$

where MLP_2 is a two-layer multi-layer perceptron (MLP) with the activation function as tanh (the last activation function is set as none). $softmax$ is performed over c to obtain the attention value $\hat{a}$, where $c = \{c_0, \ldots\}$. h^{att} is the weighted sum of $\widehat{hd}$. We concatenate h^{att} and $\overline{hd}$ to get the final utterance-level representation h. The logits of intents and slots, which also provide the unary potential value of our graph-based CRF, can be acquired by following equations:

$$logit_k^{slot} = W_{slot}\widehat{hd}_k, \tag{11}$$
$$logit^{intent} = W_{intent}h, \tag{12}$$

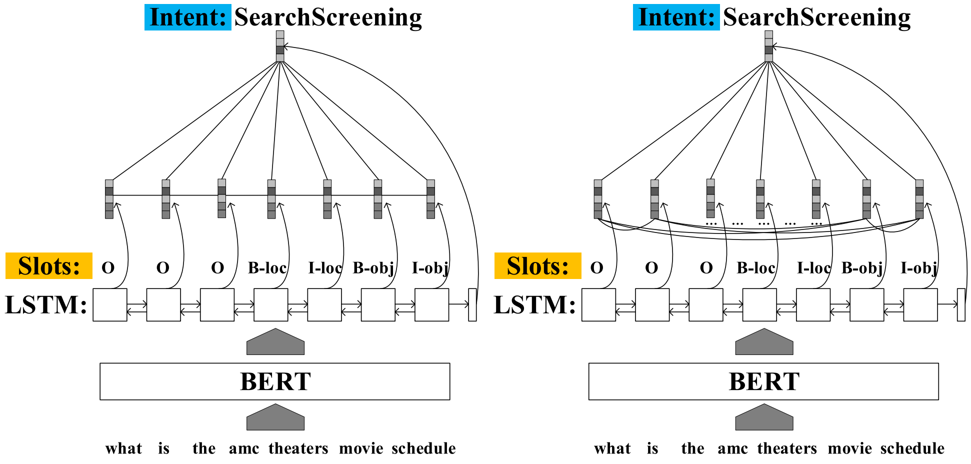
where k is the word index. $W_{slot} \in R^{h_s \times L_s}$ and $W_{intent} \in R^{h_s \times L_i}$ are the mapping matrices, where L_s and L_i are the tag size of slot and intent, respectively. h_s is the hidden size of BiLSTM.


Slots:
O
O
O
B-loc
I-loc
B-obj
I-obj
LSTM:
BERT
what is the amc theaters movie schedule
Slots:
O
O
O
B-loc
I-loc
B-obj
I-obj
LSTM:
BERT
what is the amc theaters movie schedule

(a) LC-CRF

(b) Fully-connected graph-based CRF (slot only).

Fig. 3. Two CRFs used for the slot filling task.



(a) Semi-connected graph-based CRF (joint NLU).

(b) Fully-connected graph-based CRF (joint NLU).

Fig. 4. Two CRFs used for the joint slot filling and intent detection task.

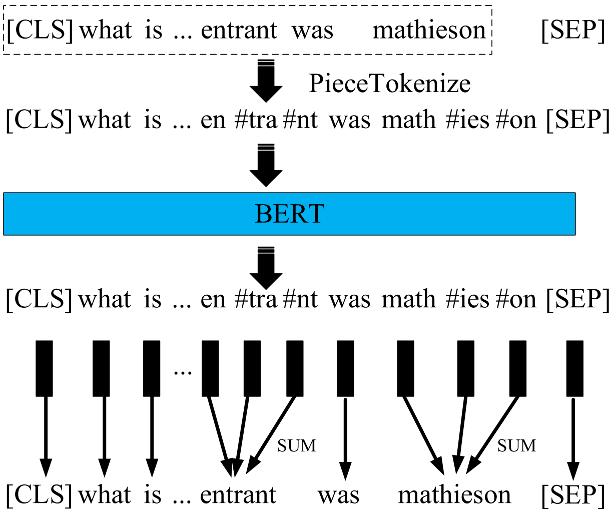

[CLS] what is ... entrant was mathieson [SEP]
PieceTokenize
[CLS] what is ... en #tra #nt was math #ies #on [SEP]
BERT
[CLS] what is ... en #tra #nt was math #ies #on [SEP]
...
SUM
SUM
[CLS] what is ... entrant was mathieson [SEP]

Fig. 5. A detailed demonstration of customized BERT.

4.3. Construction of CRF Graph

The graph of CRF is usually an Undirected Cyclic Graph (UCG). Assume that there are K words in the sentence, we can construct three different graphs depended on the specific task and connection style. We denote the adjacent matrix as *adj*.

Fully-connected Graph for Slot Filling: We can construct the adjacent matrix of words, and each word is regarded as the node. $m, n \in [0, K-1]$, note that nodes are not connected with themselves. K is the total number of words.

$$adj_{m,n} = \begin{cases} 0, & \text{if } m = n; \\ 1, & \text{other}. \end{cases} \tag{13}$$

Semi-connected Graph for Joint NLU: We can construct the adjacent matrix of the words and intent, where the intent is regarded as a new single node. $m, n \in [0, K]$, the intent node is connected with all word nodes, while each word node is only connected with its surrounding word nodes.

$$adj_{m,n} = \begin{cases} 1, & \text{if } m = n-1 \text{ or } m = n+1; \\ 1, & \text{if } n = K \text{ or } m = K; \\ 0, & \text{other}. \end{cases} \tag{14}$$

Fully-connected Graph for Joint NLU: We can construct the adjacent matrix of the words and intent, where the intent is regarded as a new single node. $m, n \in [0, K]$, note that each node is connected with all other word nodes except for itself.

$$adj_{m,n} = \begin{cases} 0, & \text{if } m = n; \\ 1, & \text{other}. \end{cases} \tag{15}$$

4.4. E2E Masked CRF as GCN

$Tag_i^t \in R^{L_s+L_i}$ denotes the tag distribution of the i-th node of t-th iteration, where $t \in [0, T], i \in [0, K]$. Tag_K^t is the tag distribution of the intent node, we add the intent as the last node. Mask mechanism is applied to Tag_i^t to recover original tag-set of the i-th node. For example, the intent node only has intent-specific tag-set with the size L_i, so the mask for intent node is $[-inf,] \ast L_s, [0,] \ast L_i$ like follows:

$$mask_i = \begin{cases} [0,] \ast L_s, [-inf,] \ast L_i, & \text{if } i < K; \\ [-inf,] \ast L_s, [0,] \ast L_i, & \text{if } i = K. \end{cases} \tag{16}$$

where L_s is the tag size of slots and L_i is the tag size of intent. T is the total number of iterations. We can obtain the unary potential by following equations:

$$una_i = \begin{cases} \left[logit_i^{slot}, Zero_{L_i}\right], & \text{if } i \in [0, K-1]; \\ \left[Zero_{L_s}, logit^{intent}\right], & \text{if } i = K. \end{cases} \tag{17}$$

where una_i is the unary potential of the i-th node. K is the total number of words. $Zero_l$ means the zero vector with the length as l. According to the mask, we need to pad the original logits by *Zero* vector.

As shown in Fig. 6, inspired by CRFasRNN [32] which implements specialized CNN to replace traditional computation procedure of CRF, we design a multi-step specialized graph convolutional network (GCN) [33,34] to conduct our graph-based CRF.

Binary Message Passing: Tag_i^0 is initialized by applying softmax to una_i. Message passing is performed by gathering and summing the current tag distribution of adjacent nodes. $adj(i)$ is the indexes of the surrounding nodes of the i-th node, which can be obtained from adj. μ is a Gaussian filter value which ranges from zero to one, and it gets lower when distance increases. α is set to 0.1 in our experiments. $\widehat{Tag_i^t}$ represents the message from surrounding nodes of the i-th node as follows:

$$Tag_i^0 = softmax(mask_i + una_i), \tag{18}$$

$$\mu(i,j) = \begin{cases} exp(-\alpha), & \text{if} i = K or j = K; \\ exp\left(-\alpha|i-j|^2\right), & \text{if} other. \end{cases} , \tag{19}$$

$$\widehat{Tag_i^t} = \sum_{j \in adj(i)} \mu(i,j) Tag_j^t. \tag{20}$$

Compatibility Transform: $W^{crf} \in R^{(L_s+L_i)\times(L_s+L_i)}$ is a transform matrix reflecting the compatibility between two specific tags. For example, $W_{i,j}^{crf}$ denotes the compatibility factor transforming from tag i to tag j:

$$\overline{Tag_i^t} = \widehat{Tag_i^t} W^{crf} + b^{crf}. \tag{21}$$

Add Unary Potentials, Masked and Softmax: Unary potentials una_i and binary potentials $\overline{Tag_i^t}$ are summed together directly to obtain the final potentials of the t-th iteration. Note that the weight of binary potentials is actually contained in W^{crf}. We further apply the mask mechanism and softmax function to re-norm the potentials as follows:

$$\widetilde{Tag_i^t} = \overline{Tag_i^t} + una_i, \tag{22}$$

$$Tag_i^{t+1} = softmax\left(mask_i + \widetilde{Tag_i^t}\right). \tag{23}$$

Eqs. (20)–(23) can be regarded as a single-layer GCN with specialized operations as shown in Algorithm 1, we repeat the process of (20)~(23) for T times to make the potentials converge.

Algorithm 1. Mean-field in graph-based CRFs, broken down to common GCN operations.

```
$t = 0$
$Tag_i^0 = softmax(mask_i + una_i)$ for all $i$ (Initialized)
repeat
    $\widehat{Tag_i^t} = \sum_{j \in adj(i)} \mu(i,j) Tag_j^t$ (Binary Message Passing)
    $\overline{Tag_i^t} = \widehat{Tag_i^t} W^{crf} + b^{crf}$ (Compatibility Transform)
    $\widetilde{Tag_i^t} = \overline{Tag_i^t} + una_i$ (Add Unary Potentials)
    $Tag_i^{t+1} = softmax\left(mask_i + \widetilde{Tag_i^t}\right)$ (Masked and Softmax)
    $t += 1$
until convergence;
```

The Illumination of W^{crf}: $W_{i,j}^{crf}$ reflects the compatibility factor when transforming from $tagset_i$ to $tagset_j$, where $tagset$ is the tagset of current dataset. We further normalize the $W_{i,j}^{crf}$ to $[-1,1]$. $W_{i,j}^{crf}! = W_{j,i}^{crf}$, which means $tagset_i - > tagset_j$ and $tagset_j - > tagset_i$ are two different processes. $W_{i,j}^{crf} \in [-1,1]$ also indicates the appearance strength of $tagset_j$ when $tagset_i$ appears. If $W_{i,j}^{crf} > 0$, then the appearance of $tagset_i$ will promote the appearance of $tagset_j$. If $W_{i,j}^{crf} < 0$, then the appearance of $tagset_i$ will inhibit the appearance of $tagset_j$. If $W_{i,j}^{crf} = 0$, then the appearance of $tagset_i$ does not affect the appearance of $tagset_j$. $|W_{i,j}^{crf}|$ reflects the strength of promotion or inhibition. Note that if $W_{i,j}^{crf}, W_{j,i}^{crf} > 0$ and the values of $|W_{i,j}^{crf}|, |W_{j,i}^{crf}|$ are relatively high, then we call $tagset_i$ and $tagset_j$ as a mutual promotion pair (MPP). If $W_{i,j}^{crf}, W_{j,i}^{crf} < 0$ and the values of $|W_{i,j}^{crf}|, |W_{j,i}^{crf}|$ are relatively high, then we call $tagset_i$ and $tagset_j$ as a mutual inhibition pair (MIP). More detailed analysis of W^{crf} will be discussed in Experiments section.

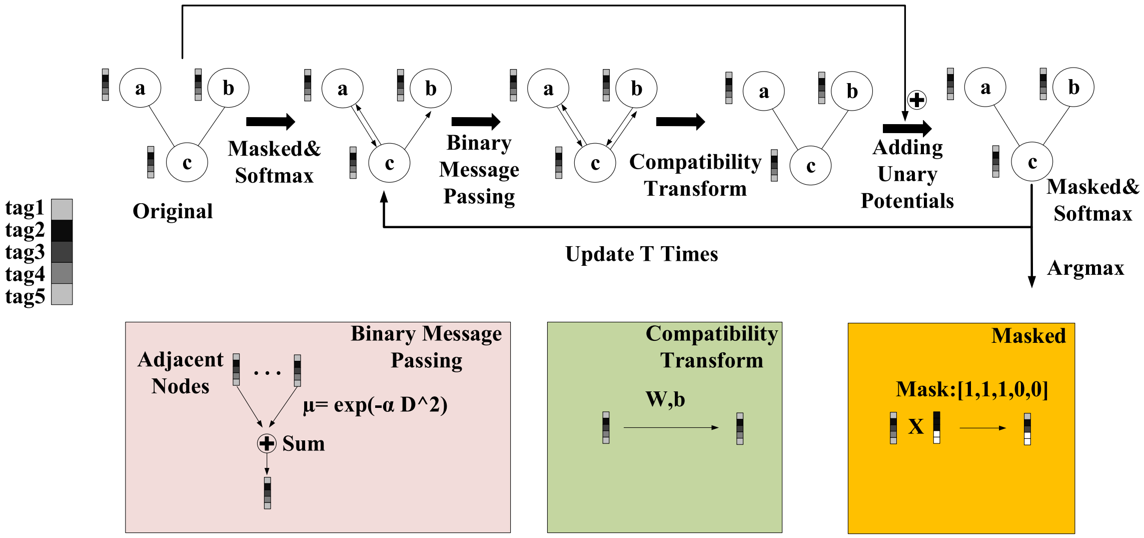


Fig. 6. A detailed demonstration of the computation procedure of an E2E graph-based CRF, which can be regarded as a multi-layer GCN.

4.5. Loss function

We recover the tag probability distribution of slots and intent by properly splitting the last output of graph-based CRF. The whole procedure of loss function is formulated as below:

$$p_i^{slot} = Tag_i^T, i \in [0, K-1], \tag{24}$$

$$p^{intent} = Tag_K^T, \tag{25}$$

$$loss = \sum_{j=0}^{K-1} - log\left(p_j^{slot}\left(y_j^{slot}\right)\right) + \lambda(-log(p^{intent}(y^{intent}))), \tag{26}$$

where y_j^{slot} is the golden slot label of the j-th word and y^{intent} is the golden intent label. λ is the weight coefficient to balance the influence of slot filling and intent detection. K is the total number of words and T is the total number of iterations. λ is set to one in our experiments.

5. Experiments

5.1. Datasets

ATIS-2: Atis-2 is the latest version of original Atis dataset [35], which is a popular and famous dataset for joint NLU and is gathered from audio recordings of people making flight reservations. Compared with original Atis, Atis-2 contains 4478, 500, 893 samples for training, development and testing set, respectively. The number of intent tags is 21 and the number of slot tags is 120.

SNIPS: Snips is a newly released dataset [36] for joint NLU, which is gathered from seven specific applications of voice assistants. Snips contains 13,084, 700, 700 samples for training, development and testing set, respectively. The number of intent tags is 7 and the number of slot tags is 72.

Facebook NLU: Facebook NLU is a newly released dataset [37] for joint NLU, which is gathered from three specific areas, i.e. *Alarm*, *Reminder* and *Weather*. Facebook NLU contains 30,521, 4181, 8621 samples for training, development and testing set, respectively. The number of intent tags is 12 and the number of slot tags is 11.

MOVIE_ENG: Movie_eng is a newly released dataset [38] for only slot filling provided by MIT, which is gathered from queries about film information. Movie_eng contains 8798, 977, 2443 samples for training, development and testing set, respectively. The number of slot tags is 25.

RESTAURANT: Restaurant is also a newly released dataset [38] for only slot filling provided by MIT, which is gathered from spoken queries related to booking restaurants. Restaurant contains 6894, 766, 1521 samples for training, development and testing set, respectively. The number of slot tags is 17.

CoNLL-2003: CoNLL-2003 is a NER dataset [39] for only slot filling provided by CoNLL, which is gathered from spoken queries related to identifying named entities. CoNLL-2003 contains 23,499, 5942, 5648 samples for training, development and testing set, respectively. The number of slot tags is 4.

Table 1
Overall performance comparison over different SOTA models for joint NLU task (Atis-2, Snips and Facebook NLU). We use two different setups of our model, including semi-connected and fully-connected graph-based CRF. LC–CRF denotes the linear chain CRF, which is widely used in NLP tasks recently. *: The original author does not release the code and point out the version of datasets either, thus we report the results we reproduce via the code from https://github.com/sz128/slot_filling_and_intent_detection_of_SLU/.

Models	Snips			Atis-2			Facebook NLU		
	F1(slot)	acc(intent)	EM	F1(slot)	acc(intent)	EM	F1(slot)	acc(intent)	EM
Tri-CNN[22]	87.1	–	–	94.4	–	–	–	–	–
Joint-Seq[4]	87.3	96.9	73.2	94.2	92.6	80.7	–	–	–
BiLSTM(Att)[23]	87.8	96.7	74.1	94.2	91.1	78.9	94.8	99.1	90.9
Slot-Gated Att[24]	88.8	97.0	75.7	94.8	93.6	82.2	95.4	99.2	91.5
Capsule-NLU[8]	91.8	97.3	80.9	95.2	95.0	83.4	95.8	99.2	91.8
BERT+Slot-Gated Att[24]	96.3	98.5	92.4	95.5	98.2	88.5	96.5	99.4	92.1
BERT+Capsule-NLU[8]	96.4	98.8	92.7	95.7	98.3	88.8	96.7	99.4	92.3
BERT+BiLSTM(Att)[10]*	96.2	98.0	92.0	95.2	97.8	87.9	96.2	99.3	91.8
+LC–CRF[10]*	96.3	98.0	92.2	95.4	97.8	88.1	96.4	99.3	91.9
+semi-E2EMG-CRF	96.3	99.2	92.9	95.5	98.6	89.0	96.5	99.6	92.2
+fully-E2EMG-CRF	97.2	99.7	93.6	96.4	99.0	89.6	97.3	99.8	92.9

Table 2
Overall performance comparison over different models for only slot filling task. We use the version of our graph-based CRF without mask (only for slot filling). Five datasets (Atis-2, Snips, Movie_eng, Restaurant and CoNLL-2003) are used, note that slot filling performance of Atis-2 and Snips with only slots will drop slightly owing to the absence of joint mechanism.

Models	Atis-2(slot only)	Snips(slot only)	Movie_eng	Restaurant	CoNLL-2003
	F1	F1	F1	F1	F1
BiLSTM[23]	93.4	87.6	82.1	72.9	83.4
DOM-Spec[38]	94.0	88.3	83.0	74.3	84.8
DOM-Spec&GEN-Adv[38]	94.2	89.2	85.3	74.5	85.2
Slot-Seq[4]	94.4	89.0	85.7	75.1	85.6
JVG[41]	94.2	88.0	82.9	73.0	–
BERT+Slot-Seq[4]	95.5	96.6	88.9	79.9	91.8
BERT+DOM-Spec&GEN-Adv[38]	95.3	96.4	88.5	79.6	91.6
BERT+BiLSTM[10]	94.8	96.0	87.8	78.6	91.4
+LC–CRF[10]	95.2	96.3	88.7	79.7	91.6
+fully-E2EG-CRF(slot only)	96.2	97.0	89.9	80.8	92.5

5.2. Experiment setup

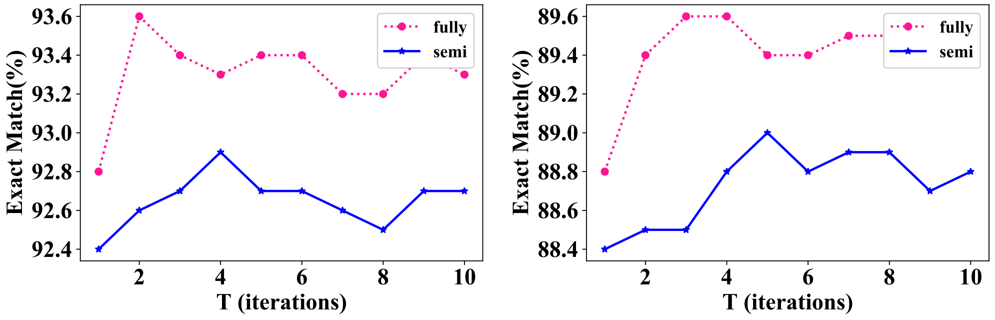
Our model takes the original utterances as input and outputs the slot label and intent label jointly. If the model is not equipped with BERT, then we use word vectors that were pre-trained on the corpora from word2vec toolkit [40]. The dimension of word embedding is set to 200. For joint NLU task, we compare our model with Tri-CNN [22], Joint-Seq [4], BiLSTM(Att) [23], Slot-Gated Att [24], Capsule-NLU [8], BERT+Slot-Gated Att [24], BERT+Capsule-NLU [8] and BERT-BiLSTM(Att) [10]. For only slot filling task, we compare our model with BiLSTM, DOM-Spec [38], DOM-Spec&GEN-Adv [38], JVG [41], Slot-Seq [4], BERT+DOM-Spec&GEN-Adv [38], BERT+Slot-Seq [4] and BERT-BiLSTM [10]. For joint NLU task, we augment BERT-Att-BiLSTM with semi-connected E2EMG-CRF and fully-connected E2EMG-CRF. For only slot filling task, we augment BERT-BiLSTM with fully-connected E2EG-CRF (mask is unnecessary). We also compare our E2E graph-based CRF with LC–CRF for all experiments.

Table 3
Overall ablation results when not using BERT for all six datasets.

Models	Atis-2	Snips	Facebook NLU	Movie_eng	Restaurant	CoNLL-2003
	EM	EM	EM	F1	F1	F1
BiLSTM	74.1	78.9	90.9	82.1	72.9	83.4
+LC–CRF	74.8	79.5	91.2	84.2	73.8	84.0
+fully-E2EG-CRF	76.9	81.6	92.3	85.8	75.5	85.2

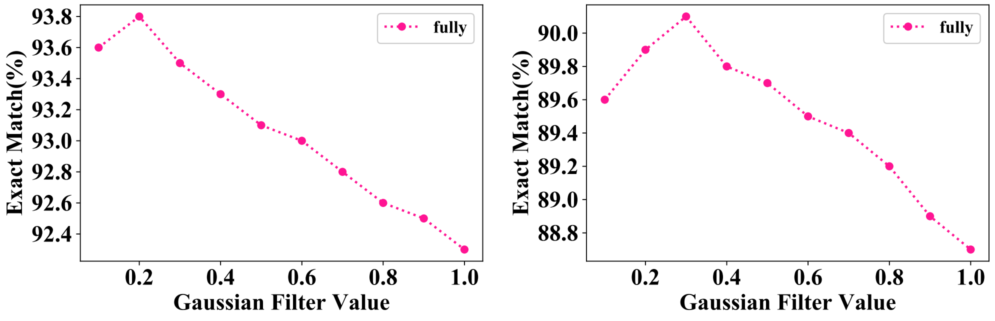
Table 4
Overall performance and time complexity comparison over different models (LC–CRF, fully-E2EMG-CRF and traditional graph-based CRF inferred via LBP algorithm) for three joint NLU tasks. All models are already based on BERT. n is the number of input words (nodes). m is the tag size. T is the number of iterations.

Models	Complexity	Snips		Atis-2		Facebook NLU	
		test speed(s/batch)	EM	test speed(s/batch)	EM	test speed(s/batch)	EM
BiLSTM(Att)	$O(nm)$	2.7	92.0	3.5	87.9	3.1	91.8
+LC–CRF	$O(nm^2)$	3.2	92.1	4.1	88.1	3.6	91.9
+Graph-based CRF(LBP)	$O(n^2m^2T)$	5.8	93.2	6.3	89.4	5.3	92.6
+fully-E2EMG-CRF	$O(nm(n+m)T)$	3.9	93.6	4.9	89.6	4.2	92.9



(a) Snips (b) Atis-2

Fig. 7. Performance over different iteration number.



(a) Snips (b) Atis-2

Fig. 8. Performance over different Gaussian filter value.

5.3. Parameter settings

Our model uses BERT-base English edition, which contains 12 hidden layers and 768 hidden units for each layer. We use Adam [42] as the optimizer for BERT and our model with the learning rate initialized by 0.00001, and decay rate of learning is set as 0.98. Except for the influence of decay rate, the learning rate decreases dynamically according to the current step number. Random sorting is applied to the training set in each training epoch. The hidden size of our basic BiLSTM is 256 and the size of all embeddings is set as 100. The vocab size of BERT is 30,522. The batch size of all models is set as 32. As for regularization, dropout function is applied to word embeddings and the dropout rate is set as 0.3. Besides, we perform L2 constraints over the parameters and L2-norm regularization is set as 0.0001. We train our model for max to 50 epochs and conduct the same experiment for 10 times. We report the average value for all metrics on testing set. Note that we use uncased version of BERT for Atis-2, Restaurant, Movie_eng, CoNLL-2003, Facebook NLU and cased version for Snips.

In next subsections, we first evaluate our model on all the four datasets. We also perform hyper-parameter study, complexity study and qualitative analysis on both datasets. At last, an error analysis is carried out for detailed performance exploration.

5.4. Metrics

To better evaluate the performance of our model, we apply and extend the modified evaluation metrics used in [43]. For our four datasets, the precision, recall, F1 [44] of slot filling and the accuracy of intent detection are reported. Exact Match (EM) is also introduced to count the testing samples with absolutely correct prediction. Precision, recall, F1 of slot filling are computed by equations below:

$$P_{slots} = \frac{pred \cap slots}{pred}, \tag{27}$$

$$R_{slots} = \frac{pred \cap slots}{slots}, \tag{28}$$

$$F1_{slots} = \frac{2P_{slots} \quad * R_{slots}}{(P_{slots} + R_{slots})}, \tag{29}$$

where *pred* is the set of the predicted slots. *slots* is the set of the golden slots. $\cap$ denotes the correction of both slot span and slot label.

5.5. Results of joint NLU and slot filling

As shown in Table 1 and the results reported in [10], BERT shows a great power when addressing sequence labeling and utterance representation problems. BERT enhanced BiLSTM has achieved the best performance among all SOTA alternatives. It is obvious that traditional LC–CRF is not able to further improve the performance according to the reported results. Semi-E2EMG-CRF shares the same connection style of words compared with LC–CRF, which results in no obvious improvement in slot F1 either. Each word is connected with intent node in semi-E2EMG-CRF, so the performance of intent accuracy and EM is still improved. For semi-E2EMG-CRF, we set the iteration number as five. Fully-E2EMG-CRF uses a fully-connected graph, in which all words are connected. Fully-E2EMG-CRF improves the EM by 1.4% and 1.5%

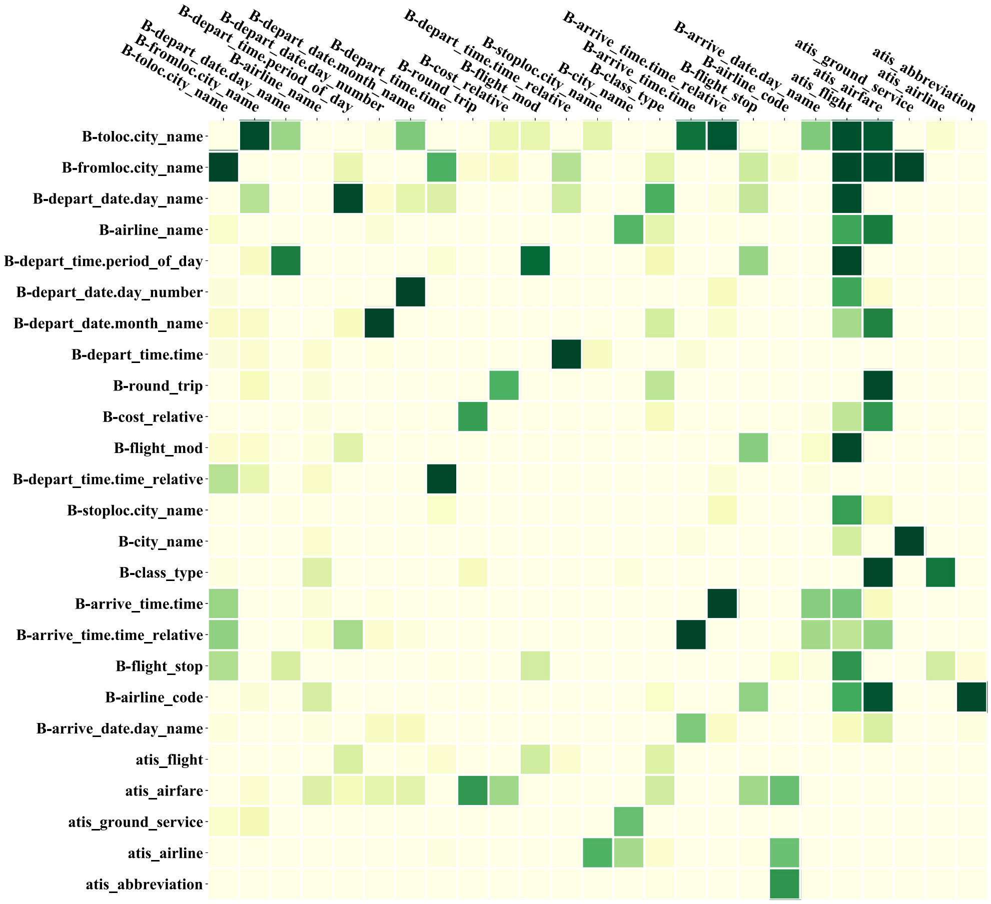


Fig. 9. Promotion transform matrix of Atis-2, darker blocks denote higher value.

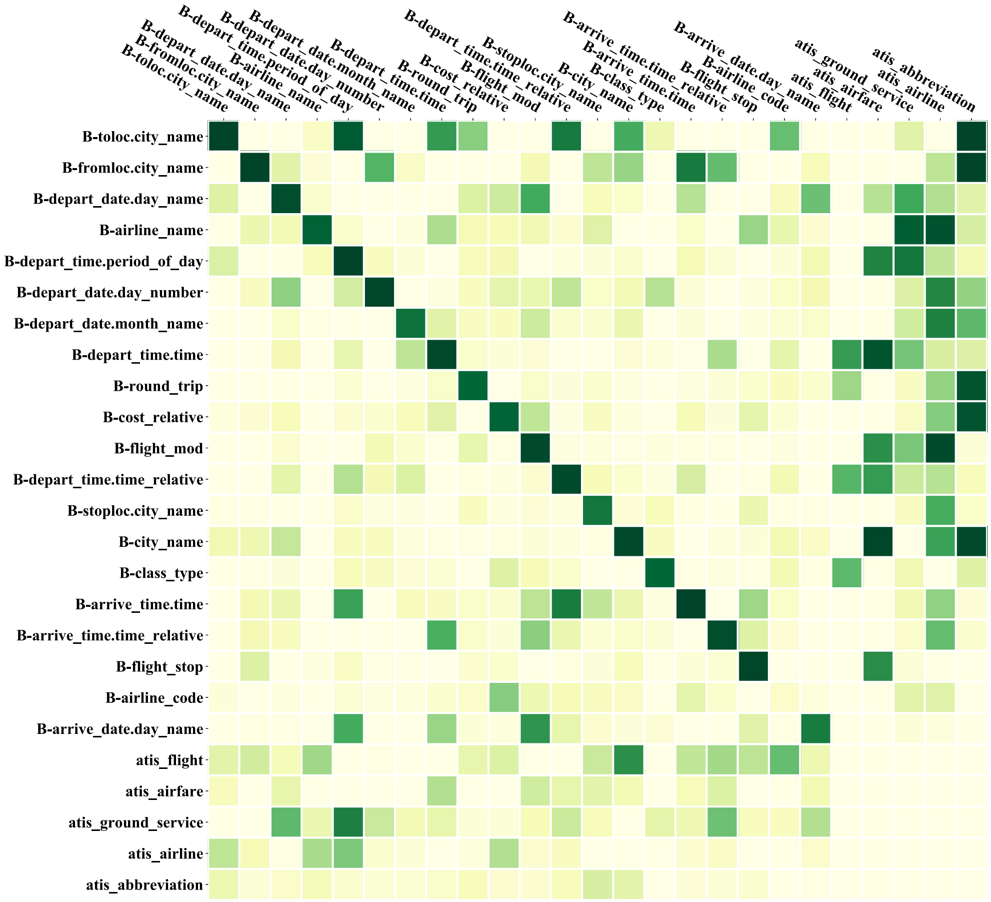


Fig. 10. Inhibition transform matrix of Atis-2, darker blocks denote higher value.

Table 5
A demonstration of two testing samples generated by LC–CRF and our model, the golden label is also included.

Model	Golden Slots	Golden Intent
text	what is meal code sb	
Golden	O O B-meal O B-meal_code	atis_abbreviation
LC–CRF	O O B-meal O B-airline_code	atis_abbreviation
E2EMG-CRF	O O B-meal O B-meal_code	atis_abbreviation
text	Put What Color Is Your Sky on the stereo	
Golden	O B-album I-album I-album I-album I-album O O B-music_item	PlayMusic
LC–CRF	O B-album I-album I-album I-album I-album O O B-music_item	AddToPlaylist
E2EMG-CRF	O B-album I-album I-album I-album I-album O O B-music_item	PlayMusic

and 1.0% for Snips, Atis-2 and Facebook NLU compared with LC–CRF. For fully-E2EMG-CRF, we set the iteration number as two. Note that all models are jointly trained for joint NLU.

As shown in Table 2, BERT still shows a great power when addressing slot filling. BERT enhanced BiLSTM has still achieved the best performance among all SOTA alternatives. Fully-E2EG-CRF (without mask and intent detection) also uses a fully-connected graph, in which all words are connected. Fully-E2EMG-CRF improves the slot F1 by 1.0% and 0.7% for Snips and Atis-2, and 1.2%, 0.9% and 1.1% for Movie_eng, CoNLL-2003 and Restaurant compared with LC–CRF. For fully-E2EG-CRF, we set the iteration number as two in all five datasets.

5.6. Ablation without BERT

As shown in Table 3, we can see that our proposed E2EG-CRF outperforms LC–CRF on all six datasets. LC–CRF indeed helps to improve the performance, but our E2EG-CRF gets better results. We can also conclude that the performance improvement between LC–CRF and E2EG-CRF in non-BERT situation is bigger than in BERT situation. We think that BERT itself improves the contextual representation, so the performance difference will shrink between LC–CRF and E2EG-CRF when applying BERT. Our method obtains better performance improvement on joint NLU tasks than slot only tasks, because slots of joint NLU have closer relationships with the surrounding slots and the utterance-level intent.

5.7. Hyper-parameter study

As shown in Table 4, we analyze the time complexity of BiLSTM, LC–CRF, fully-E2EMG-CRF and traditional Graph-based CRF(LBP), LBP algorithm requires the longest running time compared with other models. Owing to the graph structure, our fully-E2EMG-CRF costs more time than LC–CRF. Considering both the time complexity and the performance, our model is rather worthy. Actually BERT costs the most testing time in our experiment, batch is set as 128 to better assess the time interval, and our experiments is conducted on RTX-2080Ti and Core i7-8700 k.

In Fig. 7, we investigate the performance variation when utilizing different iteration value T. For fully-connected graph, we need smaller T compared with semi-connected graph, owing to the direct connection among all nodes. Semi-connected graph requires more time steps to pass the message throughout the graph. We

Table 6
A demonstration of three testing samples that our model fails, the golden label is also included.

Model	Slots	Intent
text	Play music from E-type.	
Golden	O O O B-artist O	PlayMusic
Our Model	O O O B-genre O	PlayMusic
text	how many northwest flights leave st. paul	
Golden	O O B-airline_name O O B-fromloc.city_name I-fromloc.city_name	atis_flight
Our Model	O O B-airline_name O O B-fromloc.city_name I-fromloc.city_name	atis_quantity
text	what day of the week do flights from nashville to tacoma	
Golden	O O O O O O O O B-fromloc.city_name O B-toloc.city_name	atis_day_name
Our Model	O O O O O O O O B-fromloc.city_name O B-toloc.city_name	atis_flight

thus choose $T = 2$ and $T = 5$ for fully-connected graph and semi-connected graph, respectively. When T gets too large, the performance will not be improved and the time cost becomes larger too.

In Fig. 8, we investigate the performance variation when utilizing different Gaussian filter value α which controls $exp\left(-\alpha|i-j|^2\right)$. If α is too small, then the surrounding nodes with different distance share same message passing weight. If α is too large, then the message passing weight will become too small, thus graph-based CRF is not functional in this situation. We choose $\alpha = 0.1$ for the convenience, but actually the best *alpha* is located between $0.1 \sim 0.3$ for different datasets.

5.8. Qualitative analysis

Next, we will detailedly analyze W^{crf} gathered from Atis-2 dataset. MPP denotes the tag pair in which the two tags appear together. MIP denotes the tag pair in which the two tags do not appear in the same time. We can get two associated matrices as follows:

$$W^{promotion}_{i,j} = \begin{cases} W^{crf}_{i,j}, & \text{if } W^{crf}_{i,j} > 0; \\ 0, & \text{if} other. \end{cases}, \tag{30}$$

$$W^{inhibition}_{i,j} = \begin{cases} -W^{crf}_{i,j}, & \text{if } W^{crf}_{i,j} < 0; \\ 0, & \text{if} other. \end{cases}, \tag{31}$$

$$W^{promotion}, W^{inhibition} = Norm\left(W^{promotion}\right), Norm\left(W^{inhibition}\right), \tag{32}$$

where $W^{promotion}$ is a promotion matrix, and $W^{inhibition}$ is an inhibition matrix. *Norm* is a normalization function mapping values to $[0, 1]$. $W^{promotion}_{i,j}$ reflects the promotion strength of the appearance of $tagset_j$ when $tagset_i$ appears. $W^{inhibition}_{i,j}$ reflects the inhibition strength of the appearance of $tagset_j$ when $tagset_i$ appears. We choose the most common 20 slot tags and 5 intent tags to demonstrate the partial W^{crf}.

As shown in Figs. 9 and 10, we can figure out many MPPs such as (fromloc.city_name, toloc.city_name), (round_trip, cost_relative) and (month_name, day_number). MIPs such as (period_of_day, city_name) and (period_of_day, atis_ground_service) are also obvious to search. Note that all slot tags can compose MIPs with themselves, except for 'airline_code'. We find that many samples indeed contain more than two 'airline_code' slots. We can thus conclude that most of the tags appear only once in a sample, in other words, each slot tag is unique in a single utterance.

We can still exploit other interesting phenomenons from Figs. 9 and 10. For example, when slot 'fromloc.city_name' or slot 'toloc.city_name' appear, the intent of that utterance gets little chance to be 'atis_abbreviation'. The chance of the intent to be 'atis_flight' or 'atis_airfare' becomes extremely high when slot 'airline_name' or slot 'city_name' appears. The results in these two figures demonstrate the effectiveness of our graph-based CRF owing to the interpretability of W^{crf}.

As shown in Table 5, we compare two typical testing examples annotated by LC–CRF and our E2EMG-CRF. In the first sample, two models both obtain right intent detection results, but LC–CRF annotates 'sb' as wrong slot 'airline_code'. Our model obtains the right slot 'meal_code' with the assistance of slot 'meal', owing to the direct connection among all slots in the fully-connected graph. In the second sample, LC–CRF gets right slots but wrong intent tag as 'AddToPlaylist', while our model obtains the right intent 'PlayMusic' with the assistance of slot 'music_item', owing to the direct connection between slot node and intent node.

5.9. Error analysis

As shown in Table 6, we gather three examples for which our model fails. In the first sample, our model annotates the word *'E-type'* as slot *'genre'*, actually it is the slot *'artist'*, even human beings can hardly recognize the right slot without the knowledge of *'E-type'*. In the second sample, our model obtains wrong intent tag, but the golden intent tag may be wrong because the question is obviously asked about the number of flights. In the third sample, our model outputs wrong intent tag misled by the word *'flights'* and slots about the city name, actually *'what day of the week'* is the real content.

Slot filling is more difficult than intent detection even considering the mutual influence between various slots in the same utterance. Our model achieves 93.6% and 89.6% EM on Snips and Atis-2, respectively, and most wrong prediction examples are caused by incorrect recognization of slots. Our model has almost solved the intent detection problem by obtaining 99.7% and 99.0% accuracy (intent) on Snips and Atis-2, respectively.

6. Conclusion

Traditional linear chain CRF is widely used in many NLP fields including slot filling and intent detection. Traditional LC–CRF has two obvious shortcomings including poor representation ability caused by chain structure, and limited influence range. We use graph-based CRF to handle these problems and propose a fully-connected graph-based CRF for both slot filling only tasks and joint NLU tasks. We also propose mask mechanism to adapt to multi-task or joint task, and an E2E method to accelerate the computation of CRF graph. The results demonstrate the effectiveness of the E2E graph-based CRF, and our method achieves the best performance among all published state-of-the-art models on all six datasets.

Our model can explicitly describe and utilize the mutual influence among different tag distributions. Graph structure in our model makes it possible to address multi-task problems with incompatible tag-sets and complex structure. Graph-based CRF is suitable for various kinds of classification tasks and is extremely transferable.

In future work, we will employ our method for classification tasks of other relevant domains, e.g. entity extraction, multi-label classification. We will exploit and investigate more implicit label relations in other common tasks to further testify our method.

CRediT authorship contribution statement

Hao Tang: Conceptualization, Methodology, Software, Writing - original draft. **Donghong Ji:** Writing - review & editing, Writing - review & editing, Visualization, Investigation. **Qiji Zhou:** Investigation, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the reviewers for their valuable comments. This work is supported through the grants from National Natural Science Foundation of China (NSFC-61772378), the National Key research and Development Program of China (No. 2017YFC1200500), the Research Foundation of Ministry of Education of China (No.18JZD015) and the Major Projects of the National Social Science Foundation of China (No. 11&ZD189).

References

[1] L. D. Raedt, B. Hammer, P. Hitzler, W. Maass (Eds.), Recurrent Neural Networks - Models, Capacities, and Applications, 20.01. - 25.01.2008, Vol. 08041 of Dagstuhl Seminar Proceedings, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2008, URL:http://drops.dagstuhl.de/portals/08041/..

[2] X. Luo, W. Zhou, W. Wang, Y. Zhu, J. Deng, Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data, IEEE Access 6 (2018) 5705–5715, https://doi.org/10.1109/ACCESS.2017.2785229.

[3] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[4] B. Liu, I. Lane, Attention-based recurrent neural network models for joint intent detection and slot filling, in: N. Morgan (Ed.), Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, ISCA, 2016, pp. 685–689, doi: 10.21437/Interspeech.2016-1352..

[5] S. Zhu, K. Yu, Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5–9, 2017, IEEE, 2017, pp. 5675–5679, doi: 10.1109/ICASSP.2017.7953243..

[6] X. Yang, Z. Gao, Y. Li, C. Pan, R. Yang, L. Gong, G. Yang, Bidirectional LSTM-CRF for biomedical named entity recognition, in: M. Li, N. Xiong, Z. Xiao, G. Xiao, K. Li, L. Wang (Eds.), 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2018, Huangshan, China, July 28-30, 2018, IEEE, 2018, pp. 239–242, doi: 10.1109/FSKD.2018.8687117..

[7] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1409.0473..

[8] C. Zhang, Y. Li, N. Du, W. Fan, P. S. Yu, Joint slot filling and intent detection via capsule neural networks, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 5259–5267, URL: https://www.aclweb.org/anthology/P19-1519/..

[9] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Processing 45 (11) (1997) 2673–2681, https://doi.org/10.1109/78.650093.

[10] Q. Chen, Z. Zhuo, W. Wang, BERT for joint intent classification and slot filling, CoRR abs/1902.10909, arXiv:1902.10909, URL: http://arxiv.org/abs/1902.10909..

[11] F. Tamburini, A bilstm-crf pos-tagger for italian tweets using morphological information, in: P. Basile, A. Corazza, F. Cutugno, S. Montemagni, M. Nissim, V. Patti, G. Semeraro, R. Sprugnoli (Eds.), Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5–7, 2016, Vol. 1749 of CEUR Workshop Proceedings, CEUR-WS.org, 2016, URL:http://ceur-ws.org/Vol-1749/paper_021.pdf..

[12] J. Hu, G. Wang, F. H. Lochovsky, J. Sun, Z. Chen, Understanding user's query intent with wikipedia, in: J. Quemada, G. León, Y. S. Maarek, W. Nejdl (Eds.), Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20–24, 2009, ACM, 2009, pp. 471–480, doi: 10.1145/1526709.1526773..

[13] C. Zhang, W. Fan, N. Du, P. S. Yu, Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016, ACM, 2016, pp. 1373–1384, doi: 10.1145/2872427.2874810..

[14] C. Zhang, N. Du, W. Fan, Y. Li, C. Lu, P. S. Yu, Bringing semantic structures to user intent detection in online medical queries, in: J. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, M. Toyoda (Eds.), 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11–14, 2017, IEEE Computer Society, 2017, pp. 1019–1026, doi: 10.1109/BigData.2017.8258025..

[15] Y. Chen, D. Hakkani-Tür, G. Tür, J. Gao, L. Deng, End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding, in: N. Morgan (Ed.), Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, ISCA, 2016, pp. 3245–3249, doi: 10.21437/Interspeech.2016-312..

[16] J. Gong, X. Qiu, S. Wang, X. Huang, Information aggregation via dynamic routing for sequence encoding, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018, Association for Computational Linguistics, 2018, pp. 2742–2752, URL:https://www.aclweb.org/anthology/C18-1232/..

[17] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao, S. Zhang, Investigating capsule networks with dynamic routing for text classification, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 3110–3119, URL:https://www.aclweb.org/anthology/D18-1350/..

[18] C. Xia, C. Zhang, X. Yan, Y. Chang, P. S. Yu, Zero-shot user intent detection via capsule neural networks, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 3090–3099, URL:https://www.aclweb.org/anthology/D18-1348/..

[19] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: C. E. Brodley, A. P. Danyluk (Eds.), Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, Morgan Kaufmann, 2001, pp. 282–289..

[20] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, Disan: Directional self-attention network for rnn/cnn-free language understanding, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, 2018, pp. 5446–5455, URL:https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16126..

[21] Z. Tan, M. Wang, J. Xie, Y. Chen, X. Shi, Deep semantic role labeling with self-attention, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, AAAI Press, 2018, pp. 4929–4936, URL:https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16725..

[22] P. Xu, R. Sarikaya, 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, IEEE, 2013, pp. 78–83, doi: 10.1109/ASRU.2013.6707709..

[23] D. Hakkani-Tür, G. Tür, A. Çelikyilmaz, Y. Chen, J. Gao, L. Deng, Y. Wang, Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM, in: N. Morgan (Ed.), Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016, ISCA, 2016, pp. 715–719, doi: 10.21437/Interspeech.2016-402..

[24] C. Goo, G. Gao, Y. Hsu, C. Huo, T. Chen, K. Hsu, Y. Chen, Slot-gated modeling for joint slot filling and intent prediction, in: M. A. Walker, H. Ji, A. Stent (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), Association for Computational Linguistics, 2018, pp. 753–757, URL: https://www.aclweb.org/anthology/N18-2118/..

[25] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543, URL:https://www.aclweb.org/anthology/D14-1162/..

[26] H. Caselles-Dupré, F. Lesaint, J. Royo-Letelier, Word2vec applied to recommendation: hyperparameters matter, in: S. Pera, M. D. Ekstrand, X. Amatriain, J. O'Donovan (Eds.), Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2–7, 2018, ACM, 2018, pp. 352–356, doi: 10.1145/3240323.3240377..

[27] B. Li, A. Drozd, Y. Guo, T. Liu, S. Matsuoka, X. Du, Scaling word2vec on big corpus, Data Science and Engineering 4 (2) (2019) 157–175, https://doi.org/10.1007/s41019-019-0096-6.

[28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237, URL:https://www.aclweb.org/anthology/N18-1202/..

[29] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186, URL:https://aclweb.org/anthology/papers/N/N19/N19-1423/..

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 5998–6008, URL:http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[31] B. J. Frey, D. J. C. MacKay, A revolution: Belief propagation in graphs with cycles, in: M. I. Jordan, M. J. Kearns, S. A. Solla (Eds.), Advances in Neural Information Processing Systems 10, [NIPS Conference, Denver, Colorado, USA, 1997], The MIT Press, 1997, pp. 479–485, URL:http://papers.nips.cc/paper/1467-a-revolution-belief-propagation-in-graphs-with-cycles..

[32] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr, Conditional random fields as recurrent neural networks, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015, IEEE Computer Society, 2015, pp. 1529–1537, doi: 10.1109/ICCV.2015.179..

[33] D. K. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, CoRR abs/1509.09292, arXiv:1509.09292, http://arxiv.org/abs/1509.09292..

[34] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017, URL:https://openreview.net/forum?id=SJU4ayYgl.

[35] G. Tür, D. Hakkani-Tür, L. P. Heck, What is left to be understood in atis?, in: D. Hakkani-Tür, M. Ostendorf (Eds.), 2010 IEEE Spoken Language Technology Workshop, SLT 2010, Berkeley, California, USA, December 12–15, 2010, IEEE, 2010, pp. 19–24., doi: 10.1109/SLT.2010.5700816..

[36] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, J. Dureau, Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces, CoRR abs/1805.10190, arXiv:1805.10190, http://arxiv.org/abs/1805.10190..

[37] S. Schuster, S. Gupta, R. Shah, M. Lewis, Cross-lingual transfer learning for multilingual task oriented dialog, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 3795–3805, doi: 10.18653/v1/n19-1380..

[38] B. Liu, I. Lane, Multi-domain adversarial learning for slot filling in spoken language understanding, CoRR abs/1711.11310, arXiv:1711.11310, http://arxiv.org/abs/1711.11310..

[39] J.P.C. Chiu, E. Nichols, Named entity recognition with bidirectional lstm-cnns, Trans. Assoc. Comput. Linguistics 4 (2016) 357–370, URL:https://transacl.org/ojs/index.php/tacl/article/view/792.

[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5–8, 2013, Lake Tahoe, Nevada, United States., 2013, pp. 3111–3119, URL:http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality..

[41] K. M. Yoo, Y. Shin, S. Lee, Data augmentation for spoken language understanding via joint variational generation, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 7402–7409, doi: 10.1609/aaai.v33i01.33017402..

[42] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015, URL: http://arxiv.org/abs/1412.6980..

[43] R. Xia, M. Zhang, Z. Ding, RTHN: A rnn-transformer hierarchical network for emotion cause extraction, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 5285–5291, doi: 10.24963/ijcai.2019/734..

[44] R. Wang, J. Li, Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4135–4145, URL:https://www.aclweb.org/anthology/P19-1405/..

Hao Tang is currently pursuing Ph.D. at Wuhan University, China. His current research interests are machine learning and natural language process.

Donghong Ji is Professor of School of Cyber Science and Engineering at Wuhan University. His research interests are machine learning, logics and reasoning, natural language processing and their applications in text analysis. He is co-investigator of several other externally funded projects.

Qiji Zhou is currently pursuing Ph.D. at Wuhan University, China. His current research interests are semantic parsing and natural language process.