Review of Research on Task-Oriented Spoken Language Understanding

Lixian Hou*, Yanling Li, Chengcheng Li and Min Lin

College of computer science and technology, Inner Mongolia Normal University

*cieclcc@imnu.edu.cn

Abstract. Spoken language understanding (SLU) is an important function module of the dialogue system. Slot filling and intent detection are two key sub-tasks of task-oriented spoken language understanding. In recent years, the methods of joint recognition have become the mainstream methods of spoken language understanding to solve slot filling and intent detection. Since deep neural network has advantages such as strong generalization and autonomous learning characteristics compared with traditional methods. So far, slot filling and intent detection have been developed from traditional methods to deep neural network methods, and the performance has also been significantly improved. This paper introduces the methods of tasks from the independent model to the joint model. It focuses on the joint modeling methods based on deep neural network, analyzes current problems and future development trend of two sub-tasks.

1. Introduction

Spoken language understanding is a key part of the dialogue system. The performance of spoken language understanding directly affects the entire dialogue system. A complete dialogue system is composed of five parts [1]: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialog Management (DM), Dialog Generation (DG) and Text to Speech (TTS). Spoken language understanding is divided into two tasks: Slot Filling (SF) (also known as Key Semantic Concept Recognition (KSCR) [2]) and Intent Detection (ID) [3]. Slot filling solves the problem of labeling of specific domain keywords and attributes [4], which belongs to sequential annotation task. Intent detection mainly analyzes user’s utterance behavior in input sentences, such as booking tickets and hotels, which belongs to classification problem. The keywords and intents are expressed in a semantic framework to achieve the purpose of understanding the sentence.

In recent years, deep neural network has strong generalization and may learn the features of input text automatically. It can also capture deeper semantic information in the process of learning and training. So the deep neural network method is superior to the traditional statistical machine learning method in the field of natural language processing. At the same time, due to the interdependence of slot filling and intent detection, joint recognition has become a hot research topic in spoken language understanding.

The remainder of the paper is organized as follows. In section 2, we introduce how to solve the task of slot filling and intent detection by independent model. In section 3, it elaborates that solving the slot filling and intent detection through the joint model. Section 4 concludes the paper and analyzes the future development trend of slot filling and intent detection.
2. Independent model for slot filling and intent detection

2.1. Slot filling
The slot filling is similar to the named entity recognition. It mainly extracts the key semantic components and attributes in the expression of users and identifies them with specific symbols. This task provides useful and important information for tasks such as machine translation, topic discovery, and topic tracking. Slot filling is a richer representation of named entity recognition. It is often considered a sequential annotation task.

2.1.1. Traditional methods. There are three traditional methods to solve the slot filling, including dictionary-based method [5], rule-based method [6-8] and statistics method [9-12]. At first, the dictionary-based method is relatively simple, which mainly searches for named entities in the lexicon through string matching. But there is usually no comprehensive entity library, and it is time-consuming [5]. The rule-based method mainly adds lexical rules, syntactic rules and semantic rules in the process of entity recognition. It identifies various types of named entities by rule matching. The disadvantage of the rule-based approach is that it requires domain experts and linguists to formulate rules. The mobility is relatively weak. Once new entities appear, they will conflict with the previous rules. It is time-consuming and laborious to re-formulate rules. The statistical method takes the manually annotated corpus as training set and adjusts the parameters through multiple iterating loss functions. The disadvantage is that it needs a large number of training data labeled manually and needs to construct features manually [13].

2.1.2. Deep neural network methods. In recent years, with the rapid development of deep neural networks, researchers have begun to use deep neural networks to solve slot filling tasks, such as Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), or their combination variants. Some researchers also combine deep neural network models with traditional machine learning methods.

RNN, a deep neural network model with variable length input and long distance dependence, has shown good performance in various fields, such as speech recognition, part-of-speech tagging (POS) task [14]. Although RNN has shown good performance in various fields, it has the problem of gradient disappearance. To solve this problem, the Long-Short Term Memory (LSTM) model is proposed based on RNN. For each input sequence, its context can help to obtain the grammatical and semantic information of the current word. But the directional LSTM cannot capture the relevant grammatical and semantic information of the text. Bidirectional LSTM (BLSTM) emerges as the time requires.

2.2. Intent detection

2.2.1. Traditional methods. Intent detection extracts the intent and behavior expressed in the utterance of users, which is actually a classification task. Traditional methods include Support Vector Machine (SVM) [15] (2003) and Naive Bayesian(NB) [16] (2009), Adaboost [17] (2000) and so on.

2.2.2. Deep neural network methods. Intent detection based on deep neural networks mainly uses several kinds of neural networks to classify texts. There are two main ways to input text. One is to use the average vector of all words in the discourse as the input of the neural network. The other is to use the vector of each word in the discourse as the input of the neural network [18] (2014).

Graves et al. compared the performance of the feed-forward neural network, RNN, LSTM, and Gated Recurrent Unit (GRU) in intent detection [19] (2014). Xiao et al. proposed the combination of CNN and RNN for intent detection [20] (2016).

3. The joint model for slot filling and intent detection
Generally, slot filling and intent detection are two independent tasks. In fact, the two tasks are related.
Intent detection has a positive effect on the result of slot filling and can help identify entities more accurately. So some researchers solve two tasks simultaneously. Joint model simplifies the tasks of slot filling and intent detection by training one model. Following is an explanation of the evolution of the two tasks by joint recognition. This section mainly describes the joint recognition of two tasks on ATIS [21].

3.1. The joint model of triangular-chain CRF
Because slot filling and intent are related, and traditional parallel or cascade methods cannot capture the correlation between the two tasks. Jeong et al. used triangular-chain CRF model to solve slot filling and intent detection. It captured the internal link between them [22] (2006).

Cascade mainly refers to the sequential execution of two tasks, the latter can use the results of the former as a priori knowledge to improve performance [2]. In order to compare the joint model with the cascade model, two cascade methods are adopted. The triangular-chain CRF of the research is shown in Figure 1:  $x$ represents the input text sequence, $y$ represents the corresponding output sequential annotation, $z$ represents the corresponding intent. The output of $z$ depends on two parts: one is the input text sequence, the other is the output sequential annotation $y$.

Figure 1. Triangular-chain CRF model.

Jeong's research uses a traditional statistical machine learning method CRF. Although the model contributes to joint recognition, there are still shortcomings of traditional statistical machine learning. It is time-consuming and laborious, and there must be enough training corpus.

3.2. The joint model of CNN-TriCRF
In 2013, Xu in Microsoft company used CNN-TriCRF model for solving slot filling and intent detection jointly [23]. Text features are extracted by CNN, then shared by slot filling and intent detection. TriCRF is used to normalize the whole text in slot filling task. Compared with traditional triangular-chain CRF model, this model improves the performance of intent detection and slot filling by 1% and 1.02% respectively.

Figure 2 is a joint recognition model based on CNN-TriCRF. The dotted line on the top of the graph represents the dependence between slot filling and intent detection. Feature vector $h$ generated by CNN is shared by two tasks. The model can handle multiple sequential annotation and intent classification tasks simultaneously.
Figure 2. The joint model of CNN-TriCRF.

Compared with the traditional method of extracting text features, the CNN method does not need to define features in advance. It can learn the features of input text independently. In addition, it can also solve the variable length input sequence problem. The improved CNN is used to learn text features. Finally, the learned features are used for slot filling and intent detection tasks. Formula (1) calculates of slot filling for the improved model of this research.

\[
p(Y | X) = \frac{e^{\sum_j t(Y_{i-1}, Y_i) + \sum_j h_{ij}(X, R, T) \theta_j(Y_i)}}{\sum_{Y} e^{\sum_j t(Y_{i-1}, Y_i) + \sum_j h_{ij}(X, R, T) \theta_j(Y_i)}}
\]

In the above formula, \(t(Y_{i-1}, Y_i)\) denotes the transition score from \(Y_{i-1}\) to \(Y_i\). \(h_{ij}\) denotes the \(j\) element extracted from the media window of \(X\). \(\theta_j\) denotes the feature weight corresponding to the label \(Y_i\). The improved CNN uses the text features learned for the classification task, as shown in formula (2).

\[
p(Z | X) = \frac{e^{\sum_j h_{ij} \beta_j(Z)}}{\sum_Z e^{\sum_j h_{ij} \beta_j(Z)}}
\]

Xu uses CNN to learn text features, which does not depend on manual feature extraction. For slot filling, TriCRF is used to consider the global distribution of data, instead of normalizing the data locally, so as to solve the local annotation bias problem and get the global optimal solution. But compared with traditional methods, the training parameters of the model are more and the model is more complex.

3.3. The joint model of Gated Recurrent Unit

After RecNN, Xiaodong from the Institute of Computational Language of Peking University proposed to use GRU and CNN to solve the task of slot filling and intent detection jointly [24] (2016). Through GRU, each time step is learned to predict the label of slot filling. Meanwhile, the maximum pooling layer is used to capture the global features of sentences intent classification. The model is shared by two tasks. The model of GRU-CNN is shown in Figure 3.
Figure 3. GRU-CNN model.

It used two data sets, ATIS, and CQUD (Chinese Question Understanding Dataset) collected from Baidu Know. The labeling process was manual annotated.

Although the performance of the model has been improved, there is still room for improvement. For intent detection task, the model uses the maximum pooling to classify intent, which will result in information loss. It should achieve intent detection by acquiring higher-level grammatical and semantic information.

3.4. The joint model of BLSTM-attention
Bing Liu used attention mechanism to solve slot filling and intent detection jointly [25] (2016). It used BLSTM model with attention mechanism. Attention parameter can capture additional dependence information which hidden state $h_i$ cannot capture. BLSTM-attention model is shown in Figure 4. The hidden state $h_i$ obtained by BLSTM and attention parameters $c_i$ obtained all $h_i$ are used as input for two tasks.

Figure 4. BLSTM-attention model.
BLSTM is used as the basic unit and attention parameters are added on the hidden layer state of BLSTM. The calculation of the attention parameter is as shown in the equation (3) - (5).

\[ c_i = \sum_{j=1}^{T} a_{ij} h_j \]  

\[ a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})} \]  

\[ e_{ik} = g(s_{i-1}, h_k) \]

Where \( c_i \) represents the attention vector of the \( i \)-th time step, which is weighted by all hidden layer states \( h_j \). \( a_{ij} \) denotes the weight vector, which is normalized. \( c_i \) acts on the output of the decoder of the \( i \)-th time step. The above process realizes the transformation from the initial hidden layer to the new attention layer. The weight coefficients of the hidden layer at each time reflect the influence on the current output.

The advantage of BLSTM-attention model is that it can pay attention to the effect of different input sequences on output by adding attention mechanism. But for intent detection, information entropy will be lost if we simply take the maximum value for intent detection. For slot filling, because the global normalization of the output sequence is not taken into account, it can occur bias problem easily.

3.5. Slot-gated joint recognition model based on attention

BLSTM-attention model has achieved good performance in slot filling and intent detection tasks, but it does not clearly explain how slot and intent interact. Because slot filling and intent detection depend on each other, intent can help to predict the content of slot filling to a large extent. The slot-gated mechanism is proposed by CW Goo et al. to obtain a better semantic framework for slot filling and intent detection [26] (2018). Figure 5 is the slot-gated model based on the attention.

![Figure 5. Attention-based slot-gated model.](image)

Unlike the BLSTM-attention model in section 3.4, when annotating the output of the slot sequence, the intentional attention parameter of the text is added to the output of each slot filling. It contains the...
semantic information of the whole input text sequence. It can alleviate the problem of the local optimal solution for the slot filling. It proves that slot filling and the intent detection are interdependent. Slot-gated is shown in Figure 6. Calculation such as formula (6) - (7).

\[ g = \sum v \cdot \tanh(c^s_i + W \cdot c^i) \]  \hspace{1cm} (6)
\[ y^s_i = \text{soft max}(W_o^s(h_i + c^s_i \cdot g)) \]  \hspace{1cm} (7)

Where \( c^s_i \) represents attention context vectors of the semantic slot, \( c^i \) is the attention context vectors of the intent. \( c^i \) and \( c^s_i \) have same dimension, \( v \) and \( W \) are trainable vectors and matrices respectively. Then the sum of each time series is calculated, \( g \) is the weighted feature representation of the sum. \( y^s_i \) represents slot filling calculated by the weighted feature \( c^s_i \) and \( c^i \).

Table 1 is the performance of the previous joint model for slot filling and intent detection tasks. It can be seen that the performance of joint recognition has been improving in recent years. From table 1 can be concluded that BLSTM-attention model is better than other models, compared with the triangular - chain CRF increased by 1.2%. Because the focus is on the basis of the original features for different input focus to study further. The model can capture a deeper semantic grammatical information, LSTM is the input of accumulative memory training. It is unable to focus on the type of different weights to different input distribution. So BLSTM-attention model has a good effect in slot filling and intent detection. GRU-CNN in intent detection task performance is superior to other models, compared with the triangular - chain CRF increased by 4.01%, the reason is that CNN can get the more advanced features, the performance has been improved significantly.

| Model                                      | Slot(F1) | Intent(Accuracy) |
|--------------------------------------------|----------|------------------|
| Triangular-chain CRF                       | 94.42    | 93.07            |
| CNN-TriCRF                                 | 95.42    | 94.09            |
| RecNN                                      | 93.22    | 95.40            |
| RecNN+Viterbi                              | 93.96    | 95.40            |
| GRU-CNN                                    | 95.49    | 98.10            |
| Attention encoder-decoder NN(aligned inputs) | 95.62    | 94.14            |
| BLSTM-attention                            | 95.78    | 94.40            |
| Slot-gated                                 | 95.20    | 94.10            |
3.6. Summary of joint recognition
In addition to the above methods, there are other joint models to solve slot filling and intent detection. Dilek et al. [27] (2016) proposed using RNN-LSTM model to jointly solve slot filling and intent detection. In the process, multi-domain models were established by strengthening the learning of data in different fields, and the F1 value on ATIS datasets reached 94.7%. Zhang et al. [28] (2016) proposed using Bidirectional GRU (BGRU) and adding CRF at the same time. The model realized the joint recognition of slot filling and intent detection, but there are some shortcomings. First, for intent detection, there is information loss by choosing the maximum value of GRU output each time as the intent of user discourse. Second, it does not pay attention to which words in the input sequence have a greater impact on the output of the slot filling and intent detection. Weigelt et al. [29] (2018) proposed using context, slot filling, and intent detection tasks are solved jointly with hierarchical information.

To sum up, the joint model based on deep neural network is used to solve the tasks of slot filling and intent detection. It aims to extract higher-level semantic grammatical information from text, from the previous triangular chain conditional random field to BLSTM, from relying on manual feature extraction by continuous iterative training and learning features. The performance is constantly improved. At the same time attention is also added to focus on learning parameters on the different input on the basis of the original learning characteristics. When considering the local optimal solution of the slot filling task, a CRF or slot-gated mechanism is proposed to act on the sequential output of the slot filling task after learning the advanced features. The performance of the whole model is closely related, so it is very important to extract high-quality context grammar and semantic information.

4. Conclusion and prospect
The two key tasks have experienced the expansion from traditional statistical machine method to the deep neural network model, from independent model to joint recognition, from CRF, SVM, Adaboost to RNN, CNN, LSTM, GRU, and their variants, and the application of attention mechanism. Recognition of the performance is better BLSTM-attention model and GRU-CNN model, compared with the Triangular - chain conditional random field model on slot filling and intent recognition increased by 1% - 1.36% and 1.03% 5.03% respectively. The performance is constantly improving, and further improvement is needed.

There are following three aspects that need to improve. First, the slot filling and intent detection in some research on large-scale corpus dataset have better performance. However, in the small-scale corpus (e.g., Dutch, Spanish, and other data in the collection of text sequence annotation and the lack of language), the model has not been fully made use of an application [30]. Follow-up should consider using in this respect across language knowledge migration established between cross-language dictionary. It should combine a small amount of double language as input training a larger corpus, solving in addition to the English language and so on the large-scale corpus, other languages and annotation of corpus insufficient. Second, the joint recognition model based on the deep neural network model needs to learn better grammar and semantic information. After repeated iterative training, the training parameters of the model become more and more, and the model becomes more and more complex. How to find a simplified model equivalent to the current performance to replace the existing model needs to be further studied. Third, as can be seen from the existing joint model, based on the attention mechanism for the performance of the model under the same data set is better. But the model needs for all the hidden states under different time series based on the current moment of slot filling, which uses normalized calculating weight proportion. While it can cause the model is more complex. How to use reinforcement learning model training of text feature, so as to achieve the effect of the simplified model, further to do in-depth research.

Acknowledgments
This work is partially supported by the National Natural Science Foundation of China(Nos. 61562068, 11704229, 61640204, 61806103); Natural Science Foundation of Inner Mongolia (No. 2017MS0607);
Mongolian Informatization Special Support Sub-project of Inner Mongolia Nationalities Committee (MW-2014-MGYWXXH-01); Youth Innovation and entrepreneurship talents of Inner Mongolia “grassland talents” project; Inner Mongolia science and technology planning project (research and development operation of online interactive learning platform of Mongolian language MOOC and analysis and research of national education big data); Graduate Innovation Fund of Inner Mongolia Normal University (CXJJS18112).

References
[1] Chen H, Liu X, Yin D, et al. A Survey on Dialogue Systems[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(2):25-35.
[2] LI Yanling, YAN Yonghong. Research on execution strategy about statistical spoken language understanding.Journal of Frontiers of Computer Science and Technology, 2017, 11(6): 980-987.
[3] Tur G . Spoken Language Understanding: Systems for Extracting Semantic Information from Speech[D]. Morgan & Claypool, 2011.
[4] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 363-370.
[5] wang qi. Biological named entity recognition combining dictionary and machine learning [D]. Dalian University of technology, 2009.
[6] Collins M, Singer Y. Unsupervised models for named entity classification[C]//1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999: 100-110.
[7] Cucerzan S, Yarowsky D. Language independent named entity recognition combining morphological and contextual evidence[C]//1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999: 90-99.
[8] Mikheev A, Moens M, Grover C. Named entity recognition without gazetteers[C]//Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1999: 1-8.
[9] Wei L, McCallum A. Rapid development of Hindi named entity recognition using conditional random fields and feature induction[J]. Acm Transactions on Asian Language Information Processing, 2003, 2(3):290-294.
[10] Bikel D M, Miller S, Schwartz R, et al. Nymble: a High-Performance Learning Name-finder[J]. Anlp, 1998:194–201.
[11] Bikel D M, Schwartz R, Weischedel R M. An Algorithm that Learns What's in a Name[J]. Machine Learning, 1999, 34(1-3):211-231.
[12] Borthwick A, Grishman R. A maximum entropy approach to named entity recognition[D]. New York University, Graduate School of Arts and Science, 1999.
[13] Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009: 147-155.
[14] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[15] Haffner P, Tur G, Wright J H. Optimizing SVMs for complex call classification [C]//IEEE International Conference on Acoustics. CiteSeer, 2003: I-632-I-635 vol.1.
[16] Chen J, Huang H, Tian S, et al. Feature selection for text classification with Naïve Bayes[J]. Expert Systems with Applications, 2009, 36(3):5432-5435.
[17] Schapire R E, Singer Y. BoosTexter: a boosting-based system for text categorization[J]. Machine Learning, 2000,39(2-3):135-168.
[18] Xu P, Sarikaya R. Contextual domain classification in spoken language understanding systems using recurrent neural network[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014.
IEEE International Conference on. IEEE, 2014: 136-140.

[19] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]//Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 273-278.

[20] Xiao Y, Cho K. Efficient character-level document classification by combining convolution and recurrent layers[J]. arXiv preprint arXiv:1602.00367, 2016.

[21] Tur G, Hakkani-Tür D, Heck L. What is left to be understood in ATIS?[C]//Spoken Language Technology Workshop (SLT), 2010 IEEE. IEEE, 2010: 19-24.

[22] Jeong M, Lee G G. Jointly predicting dialog act and named entity for spoken language understanding[C]//Spoken Language Technology Workshop, 2006. IEEE. IEEE, 2006: 66-69.

[23] Xu P, Sarikaya R. Convolutional neural network based triangular crf for joint intent detection and slot filling[C]//Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 78-83.

[24] Zhang X, Wang H. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding[C]//IJCAI. 2016: 2993-2999.

[25] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.

[26] Goo C W, Gao G, Hsu Y K, et al. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018, 2: 753-757.

[27] Hakkani-Tür D, Tür G, Celikyilmaz A, et al. Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM[C]//Interspeech. 2016: 715-719.

[28] Zhang X, Wang H. A Joint Model of Intent Determination and Slot Filling for Spoken Language Understanding[C]//IJCAI. 2016: 2993-2999.

[29] Weigelt S, Hey T, Landhäußer M. Integrating a dialog component into a framework for spoken language understanding[C]//Proceedings of the 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering. ACM, 2018: 1-7.

[30] Feng X, Feng X, Qin B, et al. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer[C]//IJCAI. 2018: 4071-4077.