IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2020, accepted April 25, 2020, date of publication April 30, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991484

Intention Detection Based on Siamese Neural Network With Triplet Loss

FUJI REN, (Senior Member, IEEE), AND SIYUAN XUE
Faculty of Engineering, Tokushima University, Tokushima 770-8506, Japan
Corresponding author: Fuji Ren (ren@is.tokushima-u.ac.jp)

ABSTRACT Understanding the user's intention is an essential task for the spoken language understanding (SLU) module in the dialogue system, which further illustrates vital information for managing and generating future action and response. In this paper, we propose a triplet training framework based on the multiclass classification approach to conduct the training for the intention detection task. Precisely, we utilize a Siamese neural network architecture with metric learning to construct a robust and discriminative utterance feature embedding model. We modified the RMCNN model and fine-tuned BERT model as Siamese encoders to train utterance triplets from different semantic aspects. The triplet loss can effectively distinguish the details of two input data by learning a mapping from sequence utterances to a compact Euclidean space. After generating the mapping, the intention detection task can be easily implemented using standard techniques with pre-trained embeddings as feature vectors. Besides, we use the fusion strategy to enhance utterance feature representation in the downstream of intention detection task. We conduct experiments on several benchmark datasets of intention detection task: Snips dataset, ATIS dataset, Facebook multilingual task-oriented datasets, Daily Dialogue dataset, and MRDA dataset. The results illustrate that the proposed method can effectively improve the recognition performance of these datasets and achieves new state-of-the-art results on single-turn task-oriented datasets (Snips dataset, Facebook dataset), and a multi-turn dataset (Daily Dialogue dataset).

INDEX TERMS Intention detection, BERT, RMCNN, triplet loss, fusion strategy.

I. INTRODUCTION

The dialogue systems are being integrated into various devices and allow users to speak to the system directly to perform the specific task efficiently, such as Google Home [1] and Amazon Echo [2]. The spoken language understanding (SLU) module is an indispensable component in the dialogue system. A typical SLU module is designed to transform the spoken language into a specific semantic template that human language can be well-understood by the dialogue system. After that, the dialogue management module can facilitate future actions according to detection results in the SLU module. The role of the intention detection task in SLU is to discriminate the implicit intention by recognizing the intents of received utterances. The intent tag is a semantic label attached with each utterance in dialogue, which represents the user's intention and concise utterance interpretation [3]. Therefore, intention detection task is crucial to enhance the spoken language understanding performance in the dialogue system.

In our research, we study spoken language as described in written format. According to the real situation, it is challenging to study the spoken language because of some attributes of natural language. Firstly, the sparsity of semantic information and obscure slang in spoken language make the model difficult to interpret thoroughly [4]. For instance, the average length of some utterances is no more than 20 words. Secondly, the same underlying utterances have different tags or multiple tags, which give rise to ambiguity in classifying intention labels. We use the utterance 'Yeah' as an example showed in Table 1 that the 'Yeah' has three tags, which are 'Backchannel,' 'Agree,' and 'Yes/No Answer,' respectively. The prior works of multi-class classification of intention detection exploit Softmax to train an encoder on labeled training data. The learned features are optimized under the supervision of Softmax, which cannot be sufficiently distinguished because it does not consider the intra-class compactness of features. The categories prediction was only focusing on

The associate editor coordinating the review of this manuscript and approving it for publication was Julien Le Kernec.

TABLE 1. A snippet of a dialogue sample. Each utterance corresponding to an intent label and a speaker label.

Speaker	Intents	Utterance
A	**Agree**	Oh, **yeah.**
A	Yes/No Question	You never think about that, do you?
B	**Yes Answer**	**Yeah**
A	Statement Opinion	I would think it would be harder to get up than it would be.
B	Backchannel	**Yeah**

finding a decision boundary, which results in poor generalization capabilities. Inspired by these observations, we assume that the intention recognition performance can benefit from constructing the robust and discriminative feature representations of the short-length utterances. To this end, we improve the conventional method by proposing a novel triplet training framework based on multi-class classification learning.

Pre-trained language models have recently proved to be very useful and efficient in learning general language representations. For instance, the BERT model is conceptually simple and empirically powerful in enormous natural language processing tasks [5]. Inspired by the pre-trained language model learning approach and transfer learning techniques, we refer to the concept of unsupervised pre-training method with triplet loss to learn a structured space of interpretable utterance representations.

Specifically, we design a two-stage process for intent classification, which includes feature embedding learning and intention prediction. In the first stage, we develop the RMCNN model and BERT model as Siamese encoder with metric learning to obtain robust and discriminative feature embeddings by minimizing the intra-class differences. In the second stage, we fuse the features from pre-trained feature embedding models and add additional relevant information as completed feature sets to predict intention labels in the downstream task.

We summarize the contributions of this paper as follows:

(1) The proposed triplet training framework learns discriminative utterance feature by using the same weights on different inputs. The triplet loss function infers a non-linear mapping in the resulting latent space, and the inter-class sample distances are maximized based on a certain margin [6].

(2) We utilize CNN, RMCNN (Bi-GRU-MCNN), and BERT as Siamese encoders to train the utterance triplets. Precisely, the RMCNN model can generate structural information, in which the RNN model can extract the global context, and a wide range of kernels of CNN can capture the fine-grained local components of utterance. Besides, we facilitate bidirectional encoder representation from transformers on enormous unlabeled data to obtain powerful context-dependent utterance features.

(3) The triplet selection turns out to be crucial for model convergency. By considering the strong correlations between dialogue context, we propose a sequential sampling strategy to keep the intention transition traits into the triplet sampling process.

(4) In the downstream task, we predict the probability distribution of each intent label based on multi-class classification learning. We obtain utterance features by fusing the features from different pre-trained feature embedding models. Besides, we extent features with relevant information as external knowledge, such as speaker information.

The rest of the paper is organized as follows: the related research methods are introduced in Section II; Section III introduces the model framework and methodology; Section IV conducts experiments on benchmark dataset; Section V analysis the result from different aspects; Section VI concludes the whole article and outlines the future work.

II. RELATED WORK

A. INTENTION DETECTION TASK

The learning methods for the intention detection task are divided into two categories: multi-class classification and sequence labeling. The multi-class classification models are SVM [7], Naive Bayes [8], and Maximum entropy [9] in experiments. The sequence labeling methods are HMM [7] and SVM-HMM [10]. Plenty of features had been exploited in traditional models, including lexical, syntactic features, prosodic cues, and dialogue structure. For example, the keywords [11] and vocabulary pairs as lexical features [12] can highlight the particularity of a sentence. Besides, the syntactic features like utterance length [10] and word order [13] had shown its utility for identifying intention tags. However, the traditional approaches for intention detection relied on hand-crafted features that were time-consuming and labor-intensive.

The emergence of deep learning methods effectively alleviated the constraints of the traditional approaches and achieved state-of-the-art results from natural language processing to computer vision [14]. For example, Khanpour *et al.* [15] utilized the pre-trained word embedding matrix and a modified RNN model to represent the utterance features. Kim [16] used CNN as an utterance encoder with pre-trained embedding that performed well on this task. Lee and Dernoncourt [17] got the cutting edge by investigating standard RNN and CNN that incorporated preceding short texts as context to predict dialogue act tags. Besides, some researches utilized the joint learning approach to conduct the intention detection and slot filling [48], [49]. In addition, some researchers considered the contextual structure of the multi-turn dialogue, so the intention detection task also can be regarded as a sequence labeling task. Kumar *et al.* [18] utilized hierarchical Bi-LSTM to

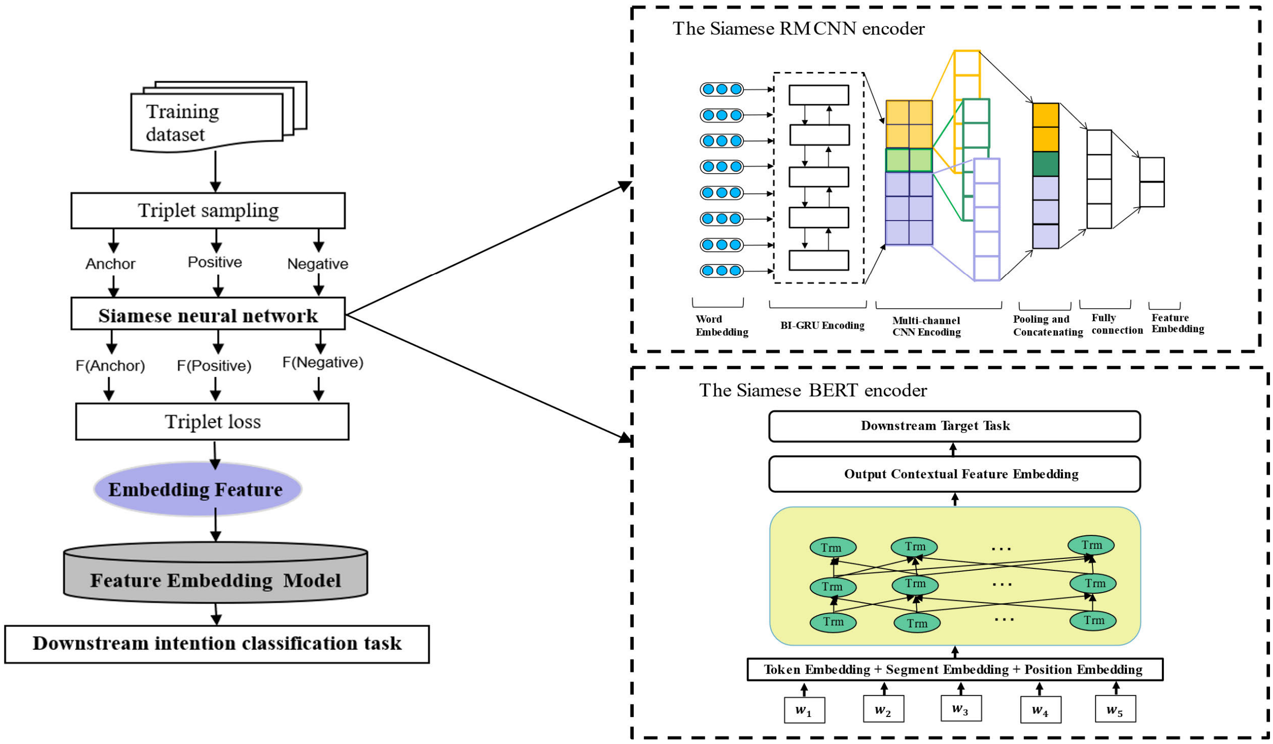

FIGURE 1. The whole intention detection framework with pre-trained feature embedding models (RMCNN, BERT).

capture utterance granularity and inherent properties from multi-levels of conversation and predicted sequential dialogue act with the CRF model. Tu *et al.* [19] build a hybrid neural network-based ensemble model for Chinese hierarchy dialogue. Notably, this paper incorporated the speaker changing as a feature to illustrate utterance peculiarity. Furthermore, some other features were useful to generate more discriminative predictions in detecting user's intention. For examples, the location of the comment in web forum [20], speaking preference of users [20], dialogue topic context of same user [21], emotion transition trait of user's blog[22], the rating and comments of products in shopping website were treated as the weak label to learn the sentence representation [34].

B. LANGUAGE REPRESENTATION MODEL

Recently, the language representation model improved significantly in many NLP tasks, such as textual entailment, semantic similarity, reading comprehension, and question answering [23]. The language representation models can provide powerful context-dependent representations by pre-training on a large scale unlabeled data, such as Contextualized Word Representations (ELMo) [24], Generative Pre-trained Transformer (GPT) [25] and Bidirectional Encoder Representations from Transformers (BERT) [5]. Besides, these models can be easily applied to different downstream tasks with minimum parameters. Therefore, we exploit the concept of pre-trained language model representation to construct a novel utterance feature embedding model in this paper.

C. METRIC LEARNING

Utilizing the deep neural network with a distance metric to learn the feature embedding had been successfully applied to many tasks, such as face recognition [26], speech recognition [27], [28] and speaker identification. For example, FaceNet [26] of Google utilized a random semi-head triplet mining approach to make up facial picture triplets, which obtained excellent performance. He *et al.* [29] achieved outstanding performance on 3D object retrieval by proposing triplet loss and center loss. Huang *et al.* [30] applied triplet loss in training to automatically recognize emotion state in spoken language. To deal with the spoken language, Cambria [31] presented a system that directly learned mapping from speech features to a compact fixed-length speaker discriminative embedding. The triplet loss function focuses on fine-grained identification and adds the measurement of the latent state, which can help model distinguish the details.

D. MULTI-SOURCE FUSION

Generally, the exceptional performance of the classification model depended on sufficiently large training corpora to a great extent. To comprehensively understand sentences, the fusion strategy can aggregate multiple sources to enriching the features and boost learning performance [31]. Majumder *et al.* [32] fused the multimodal resources like audio, video, and text for sentiment analysis. Tay *et al.* [33] generated sentence representations by using a gating mechanism to combine the sentence token features and sentiment lexicon features. Sun *et al.* [35] detected emotional elements by using a mixed model to extract sentimental objects and their tendencies from product reviews. Specifically, the multi-stream architecture is prevalent in data fusion. For example, Simonyan and Zisserman [36] designed a model with two-stream ConvNet architecture to illustrate spatial feature and temporal features, which can achieve significant performance under the condition of limited training data by the two-stream model. Inspired by these experiments, we use the fusion strategy in the downstream task to enhance the utterance feature representation.

III. PROPOSED METHOD

Before describing the proposed method in detail, we illustrate the mathematical notation for the intention detection task. In this experiment, we deal with the intention detection task based on multi-class classification learning. Suppose, we have the number of n utterance sequences $X = \{x_1, x_2, \ldots, x_n\}$ with corresponding the sequences of intents label $Y = \{y_1, y_2, \ldots, y_n\}$. Each utterance x_i of dialogue is composed of a sequence of words $x_i = \{w_1, w_2, \ldots, w_j\}$. The purpose of this paper is that given an unseen utterance x_i, we construct a model to learn the valid feature representation better and accurately predict the corresponding intent label y_i. Besides, we evaluate the proposed model on single-turn task-oriented dialogue and multi-turn conversation. It's worth noting that the multi-turn conversation contains the speaker's role information, so we supplement the role information as a feature in the downstream task. Each utterance correspond 是 to a speaker tag $C = \{c_1, c_2, \ldots, c_n\}$.

A. THE WHOLE FRAMEWORK

This section mainly introduces the whole framework of the proposed model. The entire structure consists of three parts, which are triplet sample selection, triplet training section, and the downstream task of intention classification. Firstly, the system needs a sampling strategy to generate valid triplet data (x_i^a, x_i^p, x_i^n) as training objects. One triplet sample consists of an *anchor* sample x_i^a, a *positive* sample x_i^p, and a *negative* sample x_i^n. Then, we input all the triplet samples into the Siamese encoder and train the model with a triplet loss function. The triplet training model uses the same weights on different inputs to compute variables and accomplish a better separation between two positive related samples of the same class (x_i^a, x_i^p) and one *negative* sample (x_i^n). To avoid meaningless calculation in the training process, we need to verify whether triplet samples are valid by setting up a particular margin parameter to observe Euclidean distance between embedding triplets in the test section. After the training, we can obtain a robust pre-trained feature embedding features, which can better reflect the specific characteristics of utterance. Secondly, given the well-defined feature embedding model with parameters, we exploit it mapping utterances in the downstream task. The critical components for triplet training are the Siamese model selection and triplet data composition. Therefore, the related information of essential components and modifications are illustrated in the following subsections.

B. THE TRIPLET SIAMESE NEURAL NETWORK

1) TRIPLET LOSS TRAINING

Triplet loss function is calculated on the triplet data (x_i^a, x_i^p, x_i^n), where the (x_i^a, x_i^p) are extracted from the same intention category. We obtain the negative sample (x_i^n) in different intention category from the (x_i^a, x_i^p). We exploit the feature embedding model $f_\theta(x) \in \mathbb{R}^d$ to map utterance triplets to d-dimension Euclidean space, and the distances are measured in resulting latent space.

$$D_{ap} = \| f_\theta (x_i^a) - f_\theta (x_i^p) \|_2^2 \tag{1}$$

$$D_{an} = \| f_\theta (x_i^a) - f_\theta (x_i^n) \|_2^2 \tag{2}$$

$$\forall \left(f_\theta (x_i^a), f_\theta (x_i^p), f_\theta (x_i^n)\right) \in T \tag{3}$$

The $f_\theta(\cdot)$ refers to the Siamese encoder. The $f_\theta(x_i^a), f_\theta(x_i^p), f_\theta(x_i^n)$ are outputs from the Siamese encoder. T is the set of all possible triplets in the training set. The triplet loss optimizes model by minimizing the distance between $f_\theta(x_i^a)$ and $f_\theta(x_i^p)$ and maximizing distance between $f_\theta(x_i^a)$ and $f_\theta(x_i^n)$ by at least a margin parameter $\alpha \in \mathbb{R}^+$. The triplet loss $L_{triplet}$ is illustrated as follow:

$$\sum_i^N \left[\| f_\theta (x_i^a) - f_\theta (x_i^p) \|_2^2 - \| f_\theta (x_i^a) - f_\theta (x_i^n) \|_2^2 + \alpha \right]_+ \tag{4}$$

where N stands for the number of triplets in the training set, and i denotes the i-th triplet sample. During the triplet training, generating all possible triplets can easily be satisfied but results in slower convergence. Therefore, it is vital to select valid triplet samples to improve training efficiency. The following section is about triplet sampling strategies.

2) TRIPLET SAMPLING STRATEGY

It is crucial to comply with the triplet constraint to ensure fast convergence. The constraint of triplet selection is illustrated as follow:

$$\| f_\theta (x_i^a) - f_\theta (x_i^p) \|_2^2 + \alpha < \| f_\theta (x_i^a) - f_\theta (x_i^n) \|_2^2 \tag{5}$$

Based on the constraint, we adopt two sampling strategies to extract triplets, which are random sampling strategy and sequential sampling strategy. The random sampling strategy randomly composes triplets as a training object without order. Initially, we design a generator to random sampling two different intention categories from all intention candidates N, which generates a total of $N(N-1)/2$ anchor-positive utterance pairs. For each selected anchor-positive utterance pairs, we randomly choose one of it as a negative label and another one as a positive label. Then, we randomly select an utterance from the negative label and select two utterances from the selected positive label. We combine three selected utterances as one triplet data for training. After each epoch, we repeat sampling the triplets based on batch size.

Different from the random sampling strategy, we can find that there are specific correlations among two adjacent utterances and adjacent intents in the multi-turn dialogue dataset. For example, the 'Question' tag followed by the 'Affirmative' tag is frequently appearing together, and the 'Request' tag always connects with the 'Repeat Response' tag. However, the disadvantage of the random sampling strategy is that it composes triplets without order, so it cannot take the context into triplet selection. Therefore, the encoder might learn useless context information from random order utterances. From this point of view, we keep the intention transition traits into triplet selection. To this end, we keep the original intent sequence order as *anchor* samples. Then we randomly select other utterances the same as the intention category of *anchor* samples as *positive* samples. We form negative utterance sequences with intention category that are different from the *anchor* utterances' intention category. Then, we input the triplets into Siamese encoders to train the feature embedding models. Through the sequential sampling strategy, the Siamese encoder can learn the valid context information in training. The following sections are to illustrate the Siamese neural network.

3) SIAMESE RMCNN NEURAL NETWORK

We modify the RMCNN model as a Siamese encoder to train the utterance triplets and generate a fixed-dimension representation. Firstly, we have the number of n utterances $X = \{x_1, x_2, \ldots, x_n\}$ in the dialogue. Each utterance contains variable-length word tokens $x_i = \left\{w_1, w_2, \ldots, w_j\right\}$. After triplet sampling, we obtain utterance triplet samples. For each utterance sample in triplet, we embed word tokens into vector $E = \{e_1, e_2, \ldots, e_n\}$ through a trainable embedding matrix pre-trained on enormous unlabeled data. The bidirectional GRU model encodes sequence token embedding to produce sequences of corresponding hidden vectors $H = h_1, h_2, \ldots, h_i$, which extracts the context information by concatenating the hidden states from forward and backward directions. The operation of bidirectional GRU is formulated as follows:

$$h_t^{\rightarrow} = f_{GRU}\left(h_{t+1}, e_t\right) \tag{6}$$

$$h_t^{\leftarrow} = f_{GRU}\left(h_{t-1}, e_t\right) \tag{7}$$

$$h_t = \left[h_t^{\rightarrow}, h_t^{\leftarrow}\right] \tag{8}$$

in which h_t maintains the sequence information of the utterance. Then, we feed the output from Bi-GRU layer into the CNN layer. The CNN model can capture fine-grained local features inside a multi-dimensional filed. The convolutional operation includes a filter $W_c \in \mathbb{R}$, which is utilized to a window of l continuous word vectors to produce a new feature map. A scalar feature c_i is generated from a window of words $h_{i:i+l}$ by:

$$c_i = f\left(W_c \circ h_{i:i+l} + b_c\right) \tag{9}$$

where the symbol $\circ$ indicates the dot product operation, l refers to the width of the convolutional kernel, f is a non-linear function (*ReLU*), W_c is the convolutional matrix, and b_c is a bias term. Each kernel corresponds to an utterance detector to extract specific n-gram patterns at various granularities. The kernel applied to each possible region matrix to produce a valuable feature map:

$$C = [c_1, c_2, \ldots, c_m] \tag{10}$$

in which m is the number of the channels. The pooling layer can extract local dependencies in different regions to preserve the most useful information. Then, we apply the pooling layers to capture the most valuable feature from each feature map, which includes the global maximum pooling layer and global average pooling layer. The outputs from two pooling layers are concatenated together as the local phrase feature of dialogue:

$$\hat{c} = [gmp\{c_i\}, gap\{c_i\}] \tag{11}$$

where the *'gmp'* indicates the global maximum pooling layer and the *'gap'* indicates the global average pooling layer. Then, the outputs of the pooling layers with different widths are concatenated. Finally, three fully connected layers with *'tanh'* activation are stacked together, and an L2-normalization layer is followed behind to form final utterance embedding. The Siamese RMCNN neural network optimized by minimizing the triplet loss and Adam optimizer is used during training.

4) SIAMESE BERT NEURAL NETWORK

Here is the process that we train utterance triplet samples with the Siamese BERT model. In this section, we fine-tune the pre-trained BERT model as Siamese encoder to train utterance triplet samples. Given sequence utterances $X = \{x_1, x_2, \ldots, x_n\}$, and we sample valid triplets for training. For each utterance sample in a triplet, BERT model construct token embeddings of this utterance $E = \{e_1, e_2, \ldots, e_n\}$ by concatenating the word piece embeddings, the positional embeddings, and the segment embeddings. Then, the token vectors are feed into encoder block and are encoded by stack layers. The encoder block includes multi-attention sublayers and the position-wise fully connected sublayers. The input data of the encoder block is a sequence hidden states $H = \{h_1, h_2, \ldots, h_i\}$, so the output of encoder $S = \{s_1, s_2, \ldots, s_i\}$ is illustrated as follows:

$$a_{ij}^{(k)} = Softmax\left(\left(\frac{1}{\sqrt{d_s}}\left(W_Q^{(k)} h_i\right)^T \left(W_K^{(k)} h_j\right)\right)\right) \tag{12}$$

$$s_i^{(k)} = \sum\nolimits_{v=1}^{N} a_i^{(k)} \left(w_v^{(k)} h_j\right) \tag{13}$$

$$s_i = W_O\left[s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(k)}\right] \tag{14}$$

in which k is the number of attention heads, h is the dimension of hidden states, and d_s is the parameter of scale dot-production. The W_Q, W_K, W_v and W_O indicate the model parameters. The output of the residual connection and the normalization module $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_N\}$ are denoted below:

$$\tilde{S} = LayerNorm(H + S) \tag{15}$$

The output of the position-wise fully connected sublayer $O = \{o_1, o_2, \ldots, o_N\}$ is calculated as follows:

$$o_i = W_2 ReLU\,(W_1 \tilde{s}_i + b_1) + b_2 \tag{16}$$

in which W_1, W_2, b_1 and b_2 are the model parameters. The residual connection layer and the normalization layer are followed the encoder block. The final contextual representation $\tilde{O} = \{\tilde{o}_1, \tilde{o}_2, \ldots, \tilde{o}_N\}$ is illustrated below.

$$\tilde{O} = LayerNorm(O + \tilde{S}) \tag{17}$$

We feed the final contextual representation into three fully connected layers with '*tanh*' activation and an L2-normalization layer to get final utterance token embedding. The Siamese BERT encoder is optimized by triplet loss function by end-to-end propagation, and Adam optimizer is utilized during training.

C. FEATURE FUSION IN DOWNSTREAM TASK

1) FEATURE-BASED STRATEGY

Fine-tuning the pre-trained language model can save expensive pre-computing. The pre-trained feature representation can be easily testified on many experiments with cheaper models on top of this representation [37]. Therefore, there is no need to train complex afterward. In this paper, we verify our pre-trained feature embedding model by utilizing the feature-based strategy for the downstream task. Feature-based strategy collects utterance features from the well-defined pre-trained language model to different downstream tasks.

The intention detection task in our experiment is based on the multi-class classification learning method, which can be seen in Fig. 2. The pre-trained feature embedding models (f_{RMCNN}, f_{BERT}) can form two robust utterance representations from different semantic aspects, which are denoted below.

$$U_{RMCNN} = f_{RMCNN}(x_i) \tag{18}$$
$$U_{BERT} = f_{BERT}(x_i) \tag{19}$$

Then, we feed the utterance feature U_{BERT} and U_{RMCNN} into the fully-connect layers, respectively. We use the Softmax classifier to predict the probability distribution of intention labels, which is defined as follows:

$$Q = tanh\,(W_U U + b_U) \tag{20}$$
$$\hat{y} = Softmax\left(W_Q Q + b_Q\right) \tag{21}$$

where W_U, b_U, W_Q, and b_Q are model parameters. We take cross-entropy as the loss function and Adam as an optimizer during training. The end-to-end backpropagation is employed in the training process.

2) MULTI-FEATURE FUSION STRATEGY

The multi-source fusion strategy can effectively improve the performance of natural language learning by various relevant resources [38]. Inspired by this conception, we employ a fusion strategy to accumulate semantic information of utterance from several aspects, such as utterance granularity, dialogue structure, and speaker information, which can be seen in Fig. 3. The same sentence may express different aspects concerning different aspects. To be specific, the RMCNN model can capture the global structural features of the input sentence. The BERT model remedies the limitation of the insufficient training corpora and provides more external knowledge about common utterance words. Otherwise, the participants have different roles and speaking preferences in various domains in multi-turn conversation, which also can be regarded as a distinctive feature to enhance utterance differences. We indicate speaker information in the model as 'C''. Specifically, we use numerical values to represent different speakers.

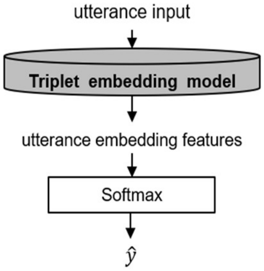

FIGURE 2. The feature-based strategy of downstream task.

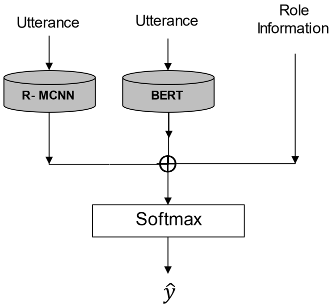

FIGURE 3. The model of fusion strategy for downstream task.

We unified a two-stream fusion model to integrate the utterance features from different models to show its different aspects. Firstly, we set two pre-trained feature embedding models as two streams to encode utterance from different aspects. We feed the sequence word tokens into the models independently and obtain the optimal parameters of each model. In this section, we compose the utterance encoder using two models with optimal settings. After the optimal parameters are trained in each stream, the outputs from each stream are concatenated together and then input to the classifier. Then, we extend the utterance representation to $U_{all} = \left[U_{RMCNN}, U_{BERT}, U_{Speaker}\right]$. Precisely, U_{RMCNN} refers to the structural feature learned from the Siamese RMCNN model, U_{BERT} refers to the fine-grained contextual feature learned from the BERT triplet model and the $U_{Speaker}$ as an additional feature refers to the speaker's role aligned with each utterance. Then, all the features are concatenated together to be a comprehensive utterance representation. The Softmax function is connected to the encoders to calculate the probability distribution, and the output is $P = \{p_1, p_2, \ldots, p_n\}$, in which n is the number of the intention labels, and p_i is the predicted probability that utterance belongs to the corresponding intent tag i, and the final predicted tag: $\hat{y} = \arg max\ (P)$. The model optimization is to minimize the cross-entropy loss, and Adam optimizer is used during training.

IV. EXPERIMENT

A. DATASETS

We evaluate the proposed model on several benchmark datasets. We find that the evaluation object of intention detection task includes not only task-oriented dialogues but also multi-turn dialogues. In the previous studies [6], the intention detection task of multi-turn conversation is regarded as a multi-class classification. Therefore, we transfer the multi-turn conversation from the nested dialogue structure into a flat structure, so that the utterance triplets can be properly sampled. Besides, we also performed a series of pre-processing steps by utilizing Stanford's CoreNLP tool [39] to avoid text noise, such as utterance tokenization and word lemmatization.

We introduce three single-turn task-oriented dialogue dataset and two multi-turn dialogue datasets, which are listed below:

The SNIPS dataset [40] is collected from the Snips personal voice assistant and contains 7 intent types. The number of samples for each intention label is approximately the same.

The ATIS dataset [41] is the audio recording of making the flight reservation. The training set includes utterances, and the test set contains 893 utterances. We follow the previous experiment and set the validation set with 500 utterances from the training set. There are 21 intention labels in the dataset.

The Facebook's multilingual dataset [42] contains annotated utterances with the English version, Spanish version, and the Thai version. It covers the weather, alarm, and reminder domains in English, Spanish, and Thai language. There are 12 intention labels in the training set.

The Daily Dialogue dataset[43] is a high-quality multi-turn dialogue dataset, which mainly records dialogue in terms of people's everyday life. Each utterance of the Daily Dialogue dataset is manually labeled with the topic tag, intention tag, and emotion tag.

The ICSI Meeting Recording Dialogue Act (MRDA) dataset [44] contains 72 hours of multi-party meeting speech dialogue from 75 naturally happened meetings. The original tag sets of MRDA included 11 general tags and 39 specific tags. Based on the previous experiments, we utilize the most widely used class-map to cluster all tags into 5 groups of intention categories.

B. HYPER-PARAMETERS TUNING

In this section, we illustrate the related parameters in model training, which is associated with the triplet training process and downstream task. All the work is implemented under the TensorFlow framework.

In terms of the triplet training with the Siamese RMCNN model, we pad each utterance to the maximum length for training. We initialized word vectors with the 300-dimensional word2vec word vectors. We set the dropout as 0.3 after the embedding layer to avoid over-fitting. The hidden size of Bi-GRU is 512 in one direction. We use multiple kernel size (1, 2, 3) in the CNN layer to encode different utterance granularity, and the filter size is 256. The three fully-connect layers and an L2-normalization layer are followed behind. We set the Adam optimizer with a learning rate of *2e-4* and a weight decay of *1e-6*.

In terms of the Siamese BERT model, we fine-tuned the BERT model with metric learning to obtain utterance features. The pre-trained BERT encoder is trained on the unlabeled data, which are Books corpus (800M words) and English Wikipedia (2500M words). The maximum length of an utterance is 50. The BERT-base model has 12-layers, 768- hidden states, and 12-heads. The hidden dim of the token embedding is 50. We set the Adam optimizer with a learning rate of *3e-5* and a weight decay of *1e-6*. The other parameters we follow the original BERT paper [5].

Furthermore, we utilize the feature-based strategy in downstream intention detection tasks. The pre-trained RMCNN and BERT feature embedding model is employed as different encoders in single-stream, respectively. In this section, we set the hidden size as 64, Adam optimizer is used with learning rate is 2e-4, and the batch size is 256.

C. BASELINES

We compare the proposed model with several state-of-the-art baseline models. For the single-turn task-oriented dataset, it includes the following:

- Attention-BiRNN [45] utilizes the encoder and decoder model for joint learning the intention detection task and slot-filling task. An attention weighted sum of all encoded hidden states is used to recognize intention.
- Slot-Gated Attention [46] uses slot-gated LSTM to learn context vector, which improves the performance of intention classification.
- Capsule-NLU [47] accomplishes the intention detection by exploiting the hierarchical semantic information. They propose a re-routing schema to synergize further the slot filling performance using the inferred intention representation.
- Joint BERT [48] uses joint intention classification and slot filling based on the pre-trained BERT model.
- BERT-SLU [49] provides a novel encoder-decoder framework based on a multi-class classification method to joint learn intention detection and slot-filling. The model uses BERT as an encoder to train utterance and then design a decoder to detect intention label.
- Cross-Lingual transfer [42] uses a novel method of using a multilingual machine translation encoder as contextual word representations to predict intents.

TABLE 2. The Dataset overviews. The number of the classes of each corpus is tag Intention, the vocabulary size of each corpus is tag Vocabulary. For the train data, validation data, and test data, we indicate the number of utterances in the table.

	Dataset	# Intention	# Vocabulary	#Train	# Validation	# Test
Single-turn	**ATIS**	21	722	4778	500	893
	Snips	7	11241	13084	700	700
	Facebook (EN)	12	3983	30521	4181	8621
	Facebook (SP)	12	1849	3617	1983	3043
	Facebook (TH)	12	1894	2156	1235	1962
Multi-turn	**DYDA**	4	25000	87170	8069	7740
	MRDA	5	10000	77900	15800	15500

TABLE 3. The recognition results on the Snips, ATIS and Facebook (EN) datasets. The evaluation criteria in this table is accuracy value of test dataset.

	Snips	ATIS	Facebook
Attention-BiRNN [45]	96.7	91.1	97.3
Slot-Gated Full-Attn [46]	96.7	93.6	93.75
Slot-Gated Intent-Attn [46]	96.8	94.1	95.43
Capsule-NLU [52]	97.3	95.0	-
Joint BERT [48]	97.3	97.5	-
Joint BERT+CRF [48]	98.6	97.9	-
BERT-SLU [49]	98.96	**99.76**	98.88
Cross-Lingual [42]	-	-	99.11
RAN-CNN	97.43	97.23	**99.13**
RAN-RMCNN	99.14	98.79	**99.12**
RAN-BERT	98.71	96.75	98.68
SEQ-CNN	98.43	98.21	**99.18**
SEQ-RMCNN	**99.29**	99.32	**99.22**
SEQ-BERT	99.00	97.31	98.97
Fusion Feature	**99.31**	99.56	**99.28**

According to previous studies, there are several multi-turn dialogue datasets contain the intention detection task. In particular, we also verify the model on the multi-turn dialogue dataset to evaluate the model generalization capability. Therefore, we compare our model with the existing baselines, which includes:

- SVM [8] is a simple baseline model, which applies the text feature and multi-classification algorithm on the dialogue act classification.
- LSTM-SoftMax [15] method applies a deep LSTM model to classify dialogue acts via the SoftMax classifier.
- CNN [17] method utilizes the CNN model to encode the utterance with the Softmax classifier. The encoder considers two preceding utterances as context information in the experiment.
- Bi-LSTM-CRF [18] method constructs a hierarchical bidirectional LSTM as an encoder to learn the conversation representation and the conditional random field as the top layer to predict intention label.
- CRF-ASN [49] incorporates hierarchical semantic inference with memory mechanism on utterance modeling at multiple levels and uses a structured attention network on the linear-chain CRF to dynamically separate the utterance into cliques.
- Dual-Attention [50] utilizes a novel dual task-specific attention mechanism to capture interaction information between intents and conversation topics for utterances.
- SelfAttn-CRF [51] proposes a hierarchical deep neural network to model different levels of utterance and dialogue act and use CRF to predict dialogue acts.

V. DISCUSSION

A. THE RESULT ANALYSIS

Table 3 and Table 4 show the intention detection accuracy on different datasets. Precisely, the prefix RAN means random triplet sampling strategy, and SEQ refers to the sequential triplet sampling strategy. The RAN-BERT means the random sampling strategy with the BERT model as Siamese encoder, and the SEQ-BERT means the sequential sampling strategy with the BERT model as a Siamese encoder. The rest model name is the same meaning.

As we can see the results shown in Table 3 and Table 4, the proposed model significantly outperforms baseline models and achieve state-of-the-art performance on Snips, Facebook (EN), and DYDA datasets. Although the proposed model does not obtain the-state-of-the-art results on ATIS and MRDA datasets, it still can show that the feature learning ability of the proposed model is useful. For the task-oriented dialogue dataset, the proposed feature learning model achieves the recognition accuracy of 99.29% (from 98.96%) on the Snips dataset, 99.22% (from 99.11%) on Facebook(EN) dataset. The fusion features also improve the performance slightly that obtain 99.31% on the Snips dataset, 99.56% on the ATIS dataset, 99.28% on Facebook(EN) dataset. For the multi-turn dialogue dataset, the model SEQ-CNN, SEQ-RCNN, and SEQ-BERT of the DYDA dataset improve the accuracy over the-state-of-the-art model by 0.6%, 2.9%, and 1.5%, respectively. The multi-source data fusion compensates for the lack of data-sparse to a certain extent. It boosts the performance than other methods because it integrates a wide range of available features, which achieves 91.3% on the DYDA dataset and 91.0% on MRDA.

However, the gains on the ATIS dataset and MRDA dataset are slight. One of the reasons for this phenomenon is that the

TABLE 4. The recognition results on the DYDA and MRDA datasets. The evaluation criteria in the table is accuracy value of test dataset.

	DYDA	MRDA
SVM [5]	75.9	82.0
LSTM-SoftMax [9]	79.6	84.6
CNN [10]	79.1	86.8
Bi-LSTM-CRF [24]	85.7	90.9
CRF-ASN [54]	-	91.7
Self-Attn-CRF [55]	-	91.1
Dual-Attn [59]	**88.1**	**92.2**
RAN-CNN	84.5	83.4
RAN-RMCNN	85.5	87.6
RAN-BERT	85.6	89.2
SEQ-CNN	**88.7**	83.6
SEQ-RMCNN	**91.0**	88.0
SEQ-BERT	**89.6**	89.6
Fusion Feature	**91.3**	91.0

data distributions in these two datasets are both imbalanced. In the MRDA dataset, the class 'Statement' is occupied more than 50% of the intention category. In the ATIS dataset, the intention label "flight" also accounts for almost half of the total training data. Based on the sampling strategy, the sampled utterances can be affected by the proportion of intent categories in the database. It is difficult for the model to learn the exact features for very few classes. Another reason is that the ambiguity of label correlation and label annotation is harmful to triplet feature learning. Besides, the MRDA dataset was found to have a high negative correlation between previous label entropy and accuracy, indicates the impact of label noise. Some utterances in ATIS dataset contains more than one label. In this experiment, we only study the single intent of utterance, which affects the results to some extent. The last reason is that the triplet training method adopts the flat dialogue structure to compose utterance triplets and predict the intents based on the multi-class classification approach in the downstream task. The model only focuses on the current utterance ignoring the hierarchical context structure information that damages the recognition performance of multi-turn conversation. In the future, we also need to consider how to be more effectively integrated triplet training with the nested structured dialogue.

B. ABLATION STUDIES

We can observe the improvement of the proposed model in the last section, and then we explore the contribution of each part in this section. We first perform ablation studies to verify the proposed feature embedding models, whether to contribute to the intention classification task. Then, we explore the details about the effect of BERT model selection. Next, we study the impact of the sampling strategy selection. Besides, the margin parameter selection also is vital for model optimization. We test the wide-range margin parameters in the experiment. Finally, we exploit the T-SNE visualization method to verify the performance of the pre-trained feature learning models.

1) THE EFFECT OF THE ENCODER SELECTION

Table 5 shows the comparison between the basic models and proposed triplet training model of different dialogue datasets. To validate the generation ability of the proposed model, we also add the other multilingual Facebook data (Spain version and Thai version) in the experiment. The CNN and RCNN models require particular text preprocessing for different languages, so there is no comparability in this experiment. Hence, we fine-tune the pre-trained multilingual BERT model to evaluate the two datasets. We implement comparative experiments under fixed hyperparameters and parameters.

The results shown in Table 5 can prove that the pre-trained feature learning models are sufficient to learn more discriminative features representation for the intention classification task. Precisely, the fine-tuned BERT model performed better than RMCNN model in basic models. However, we can see the triplet training can significantly improve the leaning ability of RMCNN. From Tabel 5, the SEQ-RMCNN model performs better than the BERT and CNN encoder on Snips datasets, ATIS dataset, Facebook dataset, and DYDA dataset. We attribute this to the fact that the combination of Wikipedia embedding and RMCNN model can effectively capture granular semantic details locally. Also, the Siamese BERT encoder improves the results of the intention classification because the pre-trained BERT model can provide rich semantic information by unsupervised trained with enormous external knowledge. The results demonstrate that the pre-trained feature embedding model can effectively improve conventional multi-class classification by supplementing utterance triplet training.

2) THE EFFECT OF THE SAMPLING STRATEGY

In this section, we discuss the effect of sampling strategy on classification results. Based on the results of Table 5, it can illustrate that both two sampling strategies can effectively improve the results of the basic models (without triplet training). To be specific, the sequential method is slightly better than the random method. Besides, the multilingual dataset also shows the sequential strategy is better than the random strategy. The SEQ-BERT improved by 0.76% over RAN-BERT in the Facebook dataset (Spain) and 2% in the Facebook dataset (Thai). The reason for these results is that the feature learning model might learn the useless context information because of random selection.

Furthermore, we make a comparison between each intention label of the DYDA dataset to show the effect of different strategies on context-sensitive data in detail. As we can see in Fig. 4, the DYDA dataset has four intention labels, which are Inform, Commissive, Question, and Directive. The proposed models generally perform great on label "Inform" and "Question" because these two intent often appears in spoken language. Although it performs poorly in tag "Commissive"

TABLE 5. The results comparison of basic model and proposed model for different dataset.

	SNIPS	ATIS	FB（EN）	FB（SP）	FB（TH）	DYDA	MRDA
CNN	97.14	96.98	98.10	-	-	79.62	81.05
RMCNN	98.57	98.77	98.13	-	-	82.14	83.54
BERT	98.63	96.62	98.42	97.08	95.80	84.21	88.05
RAN-CNN	97.43	97.23	99.13	-	-	84.56	83.47
RAN-RMCNN	99.14	98.79	99.12	-	-	85.47	87.65
RAN-BERT	98.71	96.75	98.68	96.91	94.39	85.66	89.25
SEQ-CNN	98.43	98.21	99.18	-	-	88.69	83.66
SEQ-RMCNN	**99.29**	**99.32**	**99.22**	-	-	**91.03**	88.07
SEQ-BERT	99.00	97.31	98.97	**97.67**	**96.39**	89.61	**89.69**

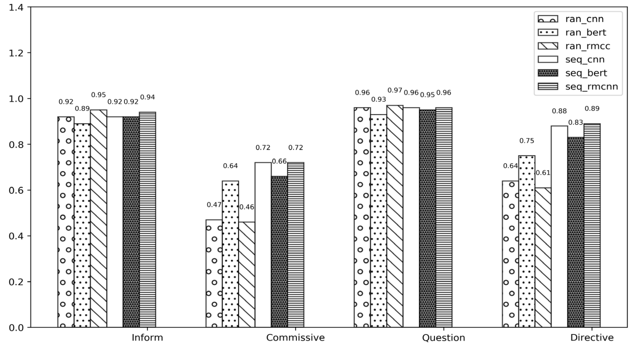

FIGURE 4. The effect of different encoders and sampling strategies on each intent in the DYDA dataset.

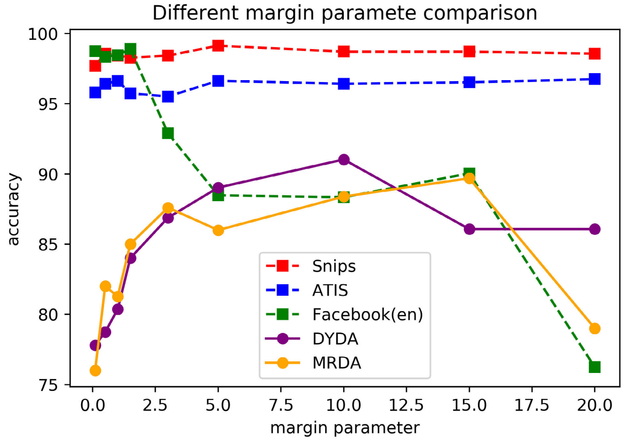

FIGURE 5. The results comparison of different margin parameter based on different dataset.

because of the lack of data, we still can find the sequential strategy can improve feature representation to be more distinguished. Specifically, the result of SEQ-CNN grew by 0.25 over RAN-CNN, the result of SEQ-RMCNN improved by 0.26 over RAN-RMCNN. The "Directive" label promotes 0.24 on CNN, 0.28 in RMCNN, only 0.08 in BERT. Therefore, the sequential sampling strategy can effectively select valid utterance triplets for spoken language objects.

3) THE EFFECT OF THE BERT MODEL SELECTION

In this section, we study the influence of the choice of the pre-trained BERT models based on the single-turn dialogue datasets. The pre-trained BERT models are publicly released on Google's GitHub website.[1] The BERT model includes a monolingual version and a multilingual version. According to the results, we find the monolingual BERT model benefits the English dataset, but it improves less on Facebook (Spain) and Facebook (Thai) datasets. The multilingual model can effectively improve the performance of the cross-language datasets. Therefore, we use monolingual models to deal with English datasets and use multilingual models to train other language datasets. Besides, the BERT models contain two uncased versions and two cased versions. Therefore, we conduct a comparison of basic BERT and BERT triplet training on the English version dataset. To keep the parameters to a minimum in the interaction system, we only verify the model on the *base* model. From Table 6, we can see the performance of uncased model is better than the cased model for utterance representation. The random sampling strategy might inferior the performance of the cased model on Snips and Facebook datasets. In the following experiments, we finally adopt the result of the Bert uncased base model as Siamese BERT encoder to train utterance triplets.

Moreover, we verified the effect of token embedding on the task-oriented dialogue dataset. We assume the token embedding might provide finer-grained semantic information of utterances compared with sentence embedding. Therefore, we facilitate the comparison between sentence embedding and token embedding on all task-oriented dialogue dataset. We indicate the T as the token embedding in Table 7 and Table 8. As we can see in Table 7 and Table 8, the token embedding can enhance the semantic information of utterance and improve the performance of intention detection. Therefore, we choose token embedding as utterance feature representation in this experiment.

4) THE EFFECT OF THE MARGIN PARAMETER

As we mentioned in (16), the margin parameter controls the relative distance between the feature embeddings to its *positive* samples and *negative* samples. Therefore, the margin parameter selection is essential for model convergency and optimization. From Fig. 5, we can observe that the triplet loss optimization is sensitive to the margin parameters. The

[1] https://github.com/google-research/bert

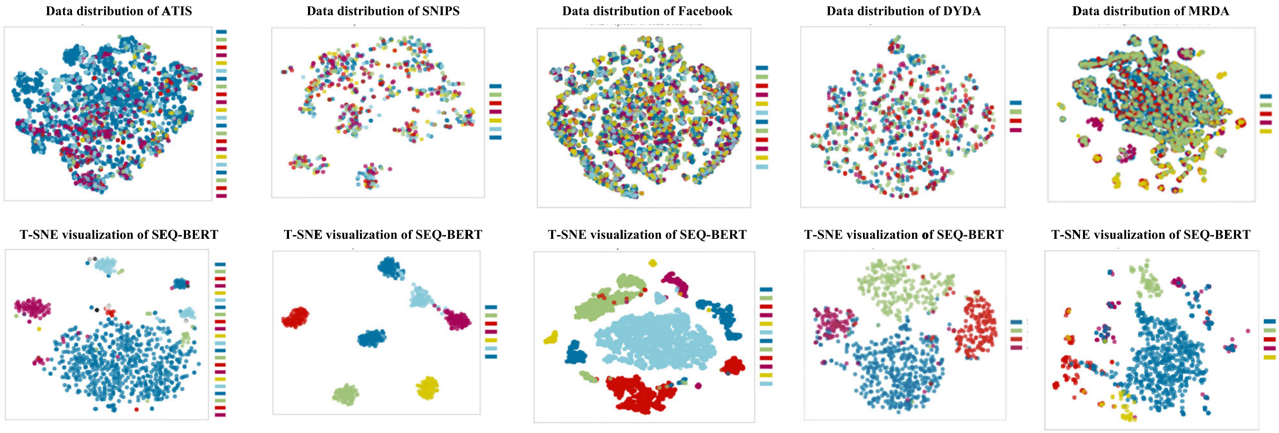

FIGURE 6. The T-SNE 2D visualization between original data distribution and pre-learned feature embeddings.

TABLE 6. The comparison of basic pre-trained BERT models and pre-trained BERT models with triplet training on ATIS, Snips, and Facebook dataset.

	Snips	ATIS	Facebook
BERT *cased base*	97.29	95.30	98.52
RAN-BERT *cased base*	96.43	95.58	98.28
SEQ-BERT *cased base*	98.14	95.63	98.53
BERT *uncased base*	98.43	95.52	98.36
RAN-BERT *uncased base*	97.86	95.75	98.63
SEQ-BERT *uncased base*	**98.97**	**97.20**	**98.90**

TABLE 7. The comparison of BERT token embedding on ATIS, Snips, and Facebook dataset.

	Snips	ATIS	FB (EN)	FB (SP)	FB (TH)
BERT	97.43	95.52	98.36	95.27	89.48
RAN-BERT	97.86	95.75	98.63	96.94	93.97
SEQ-BERT	98.97	97.20	98.90	97.47	95.15
T-BERT	98.63	96.62	98.42	97.08	95.80
T-RAN-BERT	98.71	96.75	98.68	96.91	94.39
T-SEQ-BERT	**99.00**	**97.31**	**98.97**	**97.67**	**96.39**

margin parameter is too large or too small, both results in inferior performance. The large margin parameter may cause over-fitting, and the small margin parameter may impair the strength of the triplet loss because the small value not enough to distinguish between details. Therefore, we conduct different margin parameters under fixed hyperparameters in the experiment to observe the impact of margin parameters for recognition performance. We evaluate the margin parameters on wide-ranged values from 0.1 to 20. We list the final choices of the margin parameter for each dataset. To be specific, we use 5 for the Snips dataset, 1 for the ATIS dataset, 1.5 for the Facebook dataset, and 15 for DYDA and MRDA dataset. Therefore, we set the fixed margin parameter in the following experiments.

TABLE 8. The comparison of RMCNN token embedding on ATIS, Snips, and Facebook dataset.

	Snips	ATIS	Facebook
RMCNN	97.32	96.30	97.49
RAN-RMCNN	97.42	96.58	97.88
SEQ-RMCNN	98.14	96.74	98.63
T-RMCNN	98.57	98.77	98.13
T-RAN-RMCNN	99.14	98.79	99.12
T-SEQ-RMCNN	**99.29**	**99.32**	**99.22**

5) VISUALIZATION OF LEARNED REPRESENTATION

In this section, we apply the T-SNE [52] method to visualize 2D feature embedding of test data learned from triplet learning models. Based on the T-SNE visualization method, we can intuitively observe the impacts of feature learning models on different datasets in Fig. 6. The first column is the original data distribution of each dataset, and the second column is the utterance feature embeddings of the pre-trained SEQ-BERT model. As we can see in Fig. 6, the feature embedding of the same intention category is visibly getting closer to each other and gain distinct clusters at the same time. Hence, the proposed models are benefits for extracting more discriminative features through utterance triplet training. The triplet loss training results in a better feature embedding since the margin parameter is considered appropriately.

However, the feature embedding of the MRDA corpus is not as explicit as the DYDA dataset cause the data distribution of the MRDA dataset is imbalanced. The "Statement" tags are occupied approximately 50% in test data, so the rest of the four intents are not clear enough to visualize. Therefore, this visualization reveals the intuition that better underlying feature embedding for short utterance can be obtained by Siamese neural network architecture with metric learning.

VI. CONCLUSION AND FUTURE WORK

In conclusion, we formulated the intention detection task from the perspective of enriching semantic information of utterances. In the first stage, we proposed a novel feature embedding model by utilizing the fine-tune BERT model and RMCNN model as Siamese encoders with a triplet loss function. The RMCNN and BERT as Siamese encoders were employed to train utterance triplets, and the triplet loss function can optimize the embedding model end-to-end. Then, we can obtain two well-trained feature embedding models to illustrate discriminative utterance features from different aspects. Moreover, we introduced the sequential sampling strategy in triplet selection to capture context within the dialogue. In the second stage, we used a multi-source fusion strategy to boost the recognition performance of the downstream intention detection task. Given the pre-trained models, we predict intention labels by fusing discriminative pre-trained and other relevant features within the dialogue. The extensive experiments demonstrated the effectiveness of the proposed model for intention detection on several benchmark datasets. The results illustrate that the proposed method can effectively improve the recognition accuracy of these datasets. For single-turn task-oriented dialogue, the model achieves 99.31% in the Snips dataset, 99.56% in the ATIS dataset, 99.28% in Facebook (English) dataset, 97.67% in the Facebook (Spain) and 96.39% in the Facebook (Thai). For multi-turn conversation, the recognition accuracy achieves 91.3% in the DYDA dataset and 91.0% in the MRDA dataset.

There is still much space for improvements in our system. Firstly, we can verify different neural network architectures, loss functions, and distance metrics based on the pre-training framework. Secondly, the multi-class classification learning approach may inferior the results because the model predicts intents only consider the current time step. Except for the single-turn dialogue and multi-turn dialogue, there are more complicated dialogue structures, such as multi-party and multi-modal dialogue. Therefore, the combination of intricate dialogue structures and metric learning could be a new direction. Furthermore, the triplet loss training also can be employed in other NLP tasks like emotion detection and topic adaptation in the dialogue system filed, which are also promising for future research.

REFERENCES

[1] K. Noda, "Google home: Smart speaker as environmental control unit," *Disab. Rehabil., Assistive Technol.*, vol. 13, no. 7, pp. 674–675, 2018

[2] A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor, "'Alexa is my new BFF': Social roles, user satisfaction, and personification of the Amazon echo," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, Dec. 2017, pp. 2853–2859.

[3] M. Sbisà, "Speech acts in context," *Lang. Commun.*, vol. 22, no. 4, pp. 421–436, Oct. 2002.

[4] F. Ren and K. Matsumoto, "Semi-automatic creation of youth slang corpus and its application to affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 176–189, Apr. 2016.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[6] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. 30th AAAI Conf. Artif. Intell.*, Mar. 2016, pp. 2786–2792.

[7] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguistics*, vol. 26, no. 3, pp. 339–373, Sep. 2000.

[8] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, "Dialogue act classification using a Bayesian approach," in *Proc. 9th Conf. Speech Comput.*, Saint-Petersburg, Russia, Sep. 2004, pp. 495–499.

[9] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Philadelphia, PA, USA, Mar. 2005, pp. 1061–1064.

[10] M. Tavafi, Y. Mehdad, S. Joty, G. Carenini, and R. Ng, "Dialogue act recognition in synchronous and asynchronous conversations," in *Proc. SIGDIAL*, Metz, France, Aug. 2013, pp. 117–121.

[11] M. Purver, J. Niekrasz, J. Dowding, and S. Peters, "Ontology-based discourse understanding for a persistent meeting assistant," in *Proc. AAAI Spring Symp., Persistent Assistants, Living Work (AI)*, Mar. 2005, pp. 26–33.

[12] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, California, CA, USA, Mar. 2005, pp. 1061–1064.

[13] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding MOOC discussion forums: A learning analytics approach," in *Proc. 5th Int. Conf. Learn. Anal. Knowl. (LAK)*, New York, NY, USA, 2015, pp. 146–150.

[14] N. Kalchbrenner and P. Blunsom, "Recurrent convolutional neural networks for discourse compositionality," in *Proc. Workshop Continuous Vector Space Models Compositionality*, Sofia, Bulgaria, Aug. 2013, pp. 119–126.

[15] H. Khanpour, N. Guntakandla, and R. Nielsen, "Dialogue act classification in domain-independent conversations using a deep recurrent neural network," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, Osaka, Japan, Dec. 2016, pp. 2012–2021.

[16] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[17] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2016, pp. 515–520.

[18] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, "Dialogue act sequence labeling using hierarchical encoder with CRF," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Sep. 2017, pp. 3440–3447.

[19] M. Tu, B. Wang, and X. Zhao, "Chinese dialogue intention classification based on multi-model ensemble," *IEEE Access*, vol. 7, pp. 11630–11639, Feb. 2019.

[20] Y. Jo, M. Yoder, H. Jang, and C. Rose, "Modeling dialogue acts with content word filtering and speaker preferences," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Honolulu, HI, USA, Sep. 2017, p. 2169.

[21] F. Ren and Y. Wu, "Predicting user-topic opinions in Twitter with social and topical context," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 412–424, Oct. 2013.

[22] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 5, pp. 1384–1396, Sep. 2016.

[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, Australia, vol. 1, Jul. 2018, pp. 328–339.

[24] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, vol. 1, 2018, pp. 2227–2237.

[25] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[27] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.* Cham, Switzerland: Springer, Oct. 2015, pp. 84–92.

[28] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1487–1491.

[29] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1945–1954.

[30] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech emotion recognition from variable-length inputs with triplet loss function," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 3673–3677.

[31] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[32] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[33] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Attentive gated lexicon reader with contrastive contextual co-attention for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct. 2018, pp. 3443–3453.

[34] W. Zhao, Z. Guan, L. Chen, X. He, D. Cai, B. Wang, and Q. Wang, "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, Jan. 2018.

[35] X. Sun, C. Sun, C. Quan, F. Ren, F. Tian, and K. Wang, "Fine-grained emotion analysis based on mixed model for product review," *Int. J. Netw. Distrib. Comput.*, vol. 5, no. 1, pp. 1–11, 2017.

[36] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2014, pp. 518–576.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[38] F. Chen, Z. Yuan, and Y. Huang, "Multi-source data fusion for aspect-level sentiment classification," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104831, doi: 10.1016/j.knosys.2019.07.002.

[39] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Baltimore, ML, USA, Jun. 2014, pp. 55–60.

[40] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: An embedded spoken language understanding system for private-by-design voice interfaces," 2018, *arXiv:1805.10190*. [Online]. Available: http://arxiv.org/abs/1805.10190

[41] G. Tur, D. Hakkani-Tur, and L. Heck, "What is left to be understood in ATIS?" in *Proc. IEEE Spoken Lang. Technol. Workshop*, Berkeley, CA, USA, Dec. 2010, pp. 19–24.

[42] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," 2018, *arXiv:1810.13327*. [Online]. Available: http://arxiv.org/abs/1810.13327

[43] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*. [Online]. Available: http://arxiv.org/abs/1710.03957

[44] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. 5th SIGdial Workshop Discourse Dialogue (HLT-NAACL)*, Cambridge, MA, USA, Apr. 2004, pp. 97–100.

[45] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," 2016, *arXiv:1609.01454*. [Online]. Available: http://arxiv.org/abs/1609.01454

[46] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, Jun. 2018, pp. 753–757.

[47] C. Zhang, Y. Li, N. Du, W. Fan, and P. S. Yu, "Joint slot filling and intent detection via capsule neural networks," 2018, *arXiv:1812.09471*. [Online]. Available: http://arxiv.org/abs/1812.09471

[48] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "A joint learning framework with BERT for spoken language understanding," *IEEE Access*, vol. 7, pp. 168849–168858, Nov. 2019.

[49] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, "Dialogue act recognition via CRF-attentive structured network," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 225–234, doi: 10.1145/3209978.3209997.

[50] R. Li, C. Lin, M. Collinson, X. Li, and G. Chen, "A dual-attention hierarchical recurrent neural network for dialogue act classification," 2018, *arXiv:1810.09154*. [Online]. Available: http://arxiv.org/abs/1810.09154

[51] V. Raheja and J. Tetreault, "Dialogue act classification with context-aware self-attention," 2019, *arXiv:1904.02594*. [Online]. Available: http://arxiv.org/abs/1904.02594

[52] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

FUJI REN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Japan, in 1991. From 1991 to 1994, he worked at CSK as a Chief Researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor with the Faculty of Engineering, Tokushima University. His current research interests include natural language processing, artificial intelligence, affective computing, and emotional robot. He is a fellow of the Japan Federation of Engineering Societies, IEICE, and CAAI. He is also the Editor-in-Chief of the *International Journal of Advanced Intelligence* and the Vice President of CAAI. He is the Academician of The Engineering Academy of Japan and the EU Academy of Sciences. He is also the President of the International Advanced Information Institute, Japan.

SIYUAN XUE received the B.E. degree from Capital Normal University, China, in 2015, and the master's degree from the University of Glasgow, U.K., in 2016. She is currently pursuing the Ph.D. degree with Tokushima University, Japan. Her research interests include natural language processing, dialogue systems, and deep learning.