IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 February 2024, accepted 20 February 2024, date of publication 26 February 2024, date of current version 4 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3369902

RESEARCH ARTICLE

Triple Channel Feature Fusion Few-Shot Intent Recognition With Orthogonality Constrained Multi-Head Attention

DI WU, YUYING ZHENG, AND PENG CHENG
School of Information and Electronic Engineering, Hebei University of Engineering, Handan, Hebei 056000, China

Corresponding author: Di Wu (wudiwudi@hebeu.edu.cn)

This work was supported in part by the Research Projects of the Natural Science Foundation of Hebei Province under Grant F2021402005, and in part by the National Natural Science Foundation of China under Grant 62101174.

ABSTRACT Intent recognition in few-shot scenarios is a hot research topic in natural language understanding tasks. Aiming at the problems of insufficient consideration of fine-grained features of the text and insufficient training of features in the process of model fine-tuning, the Triple Channel IntentBERT and Orthogonality Constrained Multi-Head Attention Model (TMH-IntentBERT) is proposed. The part-of-speech features, word features and keyword features are combined to extract fine-grained features of data. And the a priori knowledge of the text is fully utilized. Context information is captured through multi-head attention to learn diversified representations. At the same time, the context and score vector regularization terms are added to reduce the position and representation redundancy between heads and enhance the diversity. The experimental results show that on the public dataset, the TMH-IntentBERT model has a minimum increase of 0.63%, 0.73%, 0.79%, and 1.10% in accuracy, precision, F1 value and AUROC compared with CONVBERT, TOD-BERT, WikiHowRoBERTA, IntentBERT and DFT++, respectively.

INDEX TERMS Intention recognition, few-shot, feature fusion, multi-head attention, IntentBERT.

I. INTRODUCTION

Intelligent dialogue systems have brought a lot of convenience to users with the coming of the era of artificial intelligence [1]. Intent recognition, as an important module of spoken language understanding, is the key to the composition of human-computer dialogue systems [2]. Fine-tuning pre-trained language models on large-scale annotated datasets has been favored by many scholars in intention recognition tasks [3]. But when new domains such as neocoronary pneumonia emerge, they usually contain very few examples of data. And constructing large annotated datasets is time-consuming and laborious [4]. Therefore, studying intention recognition in few-shot scenarios becomes more important [5].

The research of intention recognition mainly includes the method based on rule template [6], machine learning [7] and deep learning [8]. Among them, the method based on deep learning is the current mainstream technology. Many researchers and scholars have conducted in-depth studies on deep learning-based few-shot intention recognition tasks. With the coming of the era of large language models (LLMs), pre-trained language models are widely used [9]. Many scholars have fine-tuned them based on BERT, and the accuracy of intention recognition has been improved largely. Compared to LLMs such as GPT, BERT requires less computational resources and memory while being easier to fine-tune and customize. Therefore, in this paper, we choose to study the BERT-based model and later use it for a few-shot intent recognition task in a task-based dialog system. Although LLM's few-shot prompting requires less data, due to its richer linguistic knowledge and representational ability, BERT will be more comprehensive and accurate when dealing with few-shot learning on a single task. Aiming at the problems of high computational cost and poor generalization ability of traditional intention recognition models, the BERT-FNN intention recognition model was proposed by Zheng and Ren [10]. Two BERT-based vectorization

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

types, word embedding and sentence embedding, were used by Kapočiūtė-Dzikienė et al. [11] to solve the problem of cross-domain intent recognition without training data. The performance of the model is improved to a great extent through the above channel fusion technique. A one-dimensional convolutional neural network and bi-directional long and short-term memory for intent recognition model was proposed by Chen et al. [12], which utilizes 1D CNN(1 Dimensional CNN - for scalar multiplications and additions) to perform convolutional operations on the target sequences, and BiLSTM to capture the dependencies at longer distances. By using standard supervised training with about 1,000 marked utterances in a public dataset, a pre-trained model called IntentBERT was proposed by Zhang et al. [13]. It can be directly applied to target domains with significantly different pre-trained data. And it is significantly better than the few-shot intent recognition of existing pre-trained models. The above methods improve the accuracy of intent recognition to a large extent. But only the sentence-level information of the text data is utilized. And there are some limitations in processing the data. The fine-grained features of the text are neglected. Aiming at existing sentiment analysis methods suffer from the problem of not being able to obtain deeper semantic connections between words, a multi-channel feature fusion sentiment analysis model based on the attention mechanism was proposed by Chen and Azragu [14]. The part-of-speech vectors, positional vectors and dependent syntactic vectors are combined into the model separately. The ability of the model to mine deeper sentiment semantic features has been improved a lot. Inspired by the above model, based on IntentBERT, lexical features and keyword features are fused in this paper. The a priori information of text data is deeply mined, and orthogonally constrained multi-attention mechanism is utilized to allocate the weights of each channel while learning contextually diverse representations.

The rest of this paper is organized as follows: Section II reviews the work related to this study. in Section III, the framework of the proposed intention recognition model TMH-IntentBERT is given, and the components of the framework are introduced in detail. In Section IV, we show the experimental research and analysis of the TMH-IntentBERT model. Section V summarizes the conclusions of this paper and future research directions.

II. RELATED WORKS

A. FEATURE FUSION

Feature fusion is combining and making full use of different features to improve the performance of the model. It has a wide range of applications in various fields such as intention recognition. A traditional machine learning approach based on multi-feature extraction was used by Qiu et al. [15] to improve the accuracy of intent classification. The feature fusion method was used by Hua and Liu [16] to enhance the data and supplement the new category samples. A dynamic multi-channel fusion mechanism was proposed by Zhou et al. [17]. GAT and BERT are used to obtain syntactic and semantic information of sentences respectively. And the multi-head attention mechanism is used to construct the connection between the two channels. A deep multi-scale feature fusion module was designed by Yu et al. [18] to interact with feature information of different scales. It has good robustness and generalization ability.

The above feature fusion method fuses different features for text tasks to improve the accuracy of text classification. However, there is a certain difference between the intention recognition task and the text classification task. The keywords in the intention recognition task are particularly important, so extracting important features is a key step in the accuracy of intention recognition. The Intentcapsnet was proposed by Xia et al. [19] to extract semantic features in discourse to distinguish existing intentions. The intentcapsnet-zsl was proposed to learn by giving zero-sample learning capability to intentcapsnet. On this basis, the feature extraction ability of IE-BERTcps-net was used by Xue and Ren [20] to extract important features. And then further the feature was used to guide the aggregation process of capsule network, which greatly improved the effectiveness of identifying unknown intentions.For single-feature text classification tasks, the quality of feature extraction affects the accuracy of classification. It is still a great challenge to improve the accuracy by fusing important features in the field of intent recognition.

To further improve the accuracy of intent recognition, many scholars have begun to study the methods of multiple feature fusion. To solve the problem that the intent recognition model does not perform well on short texts due to data sparsity, an intention recognition model was proposed by Liu and Xu [21]. Features extracted from TF-IDF, CNN and LSTM are combined to enhance short text features. And the attention mechanism is used to assign weights. Aiming at the problem of insufficient feature extraction when the current deep learning method is used to deal with the intention recognition task, a dual-channel feature fusion intention detection model was proposed by Wang et al. [22]. Pooling operation and capsule network are used to extract features to form dual-channel features. The above channel fusion technology has greatly improved the performance of the model, and the accuracy of intent recognition has been largely improved. But the text prior information is not fully utilized.

B. FEW-SHOT INTENT RECOGNITION

Intention identification is to identify the potential intention of a given utterance. To accurately identify the intention of the user, many pre-trained language models have been proposed. A dual-encoder model USE-ConveRT using a retrieval-based selection task for pre-trained was studied by Henderson et al. [23]. Zhang et al. [24] pre-trained a language model DNNC on about 100 million annotation

samples for intent recognition. Aiming at the problem of irregularities caused by short spoken texts, lack of complete context and mixed multilingualism, a mixed Chinese-English bilingual intent recognition model based on multi-feature fusion was proposed by Hu et al. [25]. The Word2vec and multilingual wordnets are combined to mask the differences between different languages. Aiming at the problem that user query terms are vague and cannot be interpreted with certainty, a hybrid deep neural network model was proposed by Xu et al. [26]. Intermediate categories are automatically generated using query logs for fine-grained query intent clarification. The performance of query intent categorization is effectively improved. The above models are mainly applied to large-scale datasets and do not take into account the application to few-shot datasets.

The intention recognition task in few-shot scenarios mainly addresses the time-consuming and laborious problem of constructing large-scale datasets. A lot of research on few-shot learning intention recognition for emerging categories has been carried out by scholars. A dynamic semantic matching and aggregation network for few-shot intent detection was proposed by Nguyen et al. [27]. Semantic components are extracted from discourse through multi-head attention and additional dynamic regularization constraints. In addition, to address the problem of slot nesting in multi-intention recognition, a multi-dimensional type-slot label interaction network was constructed by Wan et al. [28]. It enhances the correlation between intention and slot and provides more sufficient information. The above pre-trained model achieves an improvement in the effectiveness of the intent recognition task on a few-shot dataset.

C. MULTI-HEAD ATTENTION

The attention model is a standard component of deep learning networks and widely used in text categorization tasks. A syntactic-aware local attention model was proposed by Li et al. [29]. The syntactic knowledge of BERT was combined into the local attention mechanism to focus on syntactically related words. The efficiency of BERT has been improved by combining the syntactic knowledge of BERT into a local attention mechanism. The CNN was used by Ma et al. [30] to extract spatial features, and the Bi-LSTM was used to extract temporal features. The attention mechanism was used to assign weights to the two channels for weighting. Attention-based models can benefit from more focused attention on localized regions. The multi-head attention mechanism captures different situations through multiple individual attention functions. The multi-head attention mechanism is a supplement and development of the attention mechanism. To enhance the diversity among multiple attention heads, three inconsistent regularization methods were proposed by Li et al. [31]. The above methods were applied to the subspace, attention location and output representation of each attention head, respectively. Multi-head attention has position and representation redundancy, resulting in insufficient richness between heads. To solve this problem, a multi-head attention mechanism of regularization technology was proposed by Lee et al. [32]. The above attention method deeply mines the semantic relationship of the text and provides new ideas for the subsequent intention recognition task.

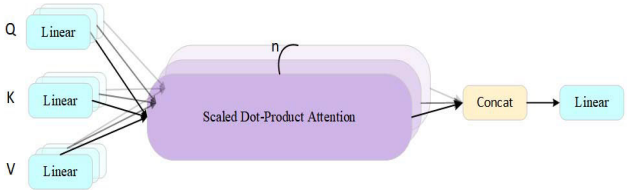

FIGURE 1. The diagram of multiple attention framework.

In recent years, the attention mechanism has been widely used in various fields. In the field of intention recognition, it has been applied to obtain context information. An attention-based convolutional neural network was proposed by Hou et al. [33] to effectively optimize the weight of global features and local features. It improves the accuracy of intent detection greatly. To enable slot semantics fully integrate intent information, a heterogeneous attention mechanism was proposed by Hao et al. [34]. A federated model was proposed by Wei et al. [35] with a wheelchart attention network that is able to directly model interrelated connections. To address the problem of intent recognition in multimodal scenes, an adaptive multimodal fusion method based on an attention-based gated neural network was designed by Huang et al. [36] to eliminate noisy features. The above attention mechanism approach improves the efficiency of the model and the accuracy of intent recognition. But fails to fully consider the problem of information redundancy between the heads of multi-head attention has not been sufficiently taken into account.

The diagram of multiple attention framework is shown in Fig.1.

In Fig.1, Q, K and V are fixed single values, followed by a linear layer. The scaled dot-attention product has n heads, which are concatenated and passed into the linear layer.

Aiming at the problem that the traditional intention recognition model fails to fully consider the text prior information and the model hidden layer information, Triple Channel IntentBERT and Orthogonality Constrained Multi-Head Attention Model (TMH-IntentBERT) is proposed. The final text representation is obtained by fusing the part-of-speech features, word features and keyword features of the text. It extracts the fine-grained features of the data and makes full use of the prior information of the text. The multi-head attention mechanism is utilized to focus on different subsequences to learn diverse representations. Context and score vector regularization terms are utilized to constrain multi-head attention to reduce position and representation redundancy. Meanwhile, the diversity between attention heads has been enhanced.

III. TRIPLE CHANNEL INTENTBERT AND ORTHOGONALITY CONSTRAINED MULTI-HEAD ATTENTION MODEL

To address the lack of priori information about the text of existing models, insufficient consideration of fine-grained features of the text and insufficient training of features in the process of model fine-tuning, the Triple Channel Intent-BERT and Orthogonality Constrained Multi-Head Attention Model(TMH-IntentBERT) is proposed. TMH-IntentBERT contains four parts: the data process layer, the feature fusion layer, the orthogonality constrained multi-head attention layer, and the fully connected layer. Firstly, in the data process layer, the text is segmented, the stop words are removed, and the case conversion is performed. Secondly, the part-of-speech features, word features and keyword features are fused in the multi-channel feature fusion layer to obtain the fusion vector. Then, the fusion vector obtained by the upper layer is input into the orthogonality constrained multi-head attention layer to obtain context information. Finally, the final intent label is obtained through the fully connected layer.

The framework of the TMH-IntentBERT model is shown in Fig.2.

A. DATA PROCESS LAYER

Word segmentation, stop words removal and case conversion on English text are performed in the data process layer. And the intent text is vectorized. The pre-trained language model BERT is used by the TMH-IntentBERT model to vectorize English text. In the word vector layer of BERT, the sum of the word vector, text vector and position vector are used as the input of the model. The vectors generated by BERT pre-trained solve the problem of multiple meanings of a word compared to static word vectors such as one-hot. And feature extraction is enhanced compared to dynamic word vectors such as ELMO. The schematic of BERT is shown in Fig.3.

Token Embeddings are used to convert words into fixed dimensions. The first word is the [CLS] for subsequent classification tasks. Segment Embeddings have only two vector representations of 0 and 1 to distinguish two sentences in a sentence pair. Position Embeddings are used to solve the problem that Transformers cannot encode the order of input sequences.

B. FEATURE FUSION LAYER

The word features of the text are obtained through the data process layer. NLTK and Rake_NLTK (Rapid Automatic Keyword Extraction algorithm) are used in the feature fusion layer to obtain the part-of-speech features and keyword features respectively. The above three features are formed into a fusion vector. Among them, to take into account textual fine-grained features, part-of-speech features are used to leverage textual a priori information. Keyword features are calculated by the frequency of words in the text to find hidden information in the text. Feature fusion can effectively improve the accuracy of intention recognition.

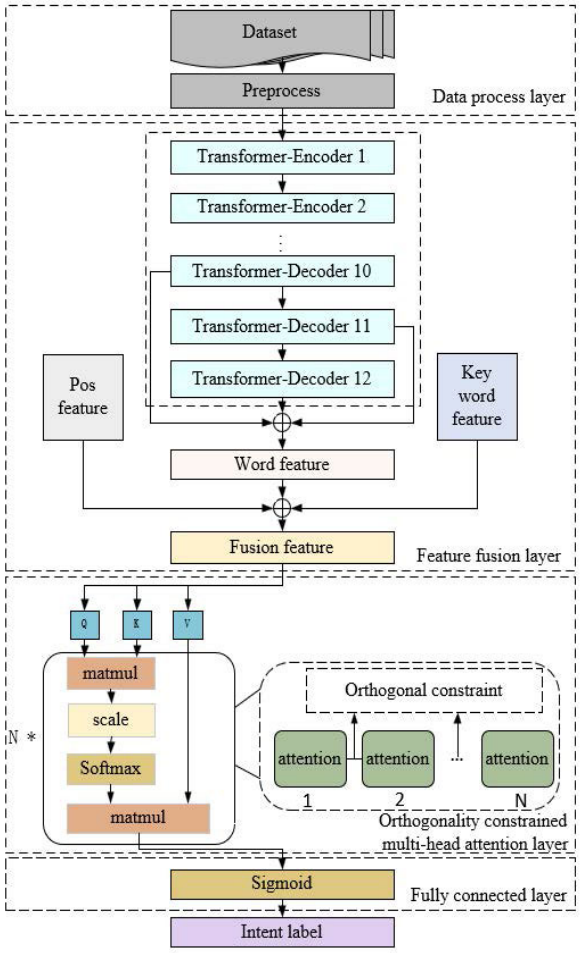

FIGURE 2. The framework of the TMH-IntentBERT model.

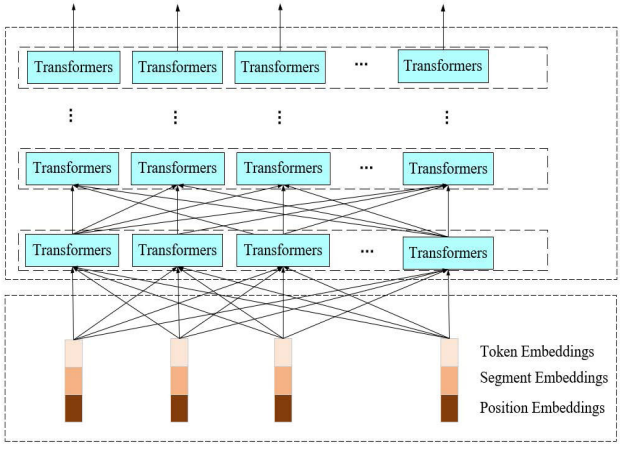

FIGURE 3. The schematic of BERT.

In the feature fusion layer, the features of the three channels of part-of-speech features, word features and keyword features are fused. By fusing the features, the hidden information in the text is fully mined, and the model intention

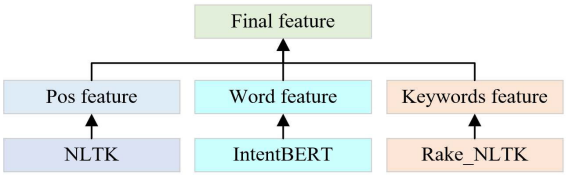

FIGURE 4. Feature fusion process.

recognition ability is improved. The feature fusion process is shown in Fig.4.

1) PART-OF-SPEECH FEATURE

The part-of-speech feature is a vector representation of the grammatical attributes of a word. Taking a single word as a unit, it is divided into different categories according to the grammatical rules of the language used and the meaning of the word itself. NLTK is used to label the original input. And then the labeled information is vectorized to form a multi-dimensional continuous value part-of-speech vector matrix. If a sentence s of length n is entered, e_n^p denotes the vectorization matrix, the equation is as follows.

$$e_n^p = tag_1 + tag_2 + tag_3 + \ldots + tag_n \tag{1}$$

where tag_n is the part-of-speech vector of the n-th word, $tag_n \in V^b$, and the dimension of the vector is b. p represents the part-of-speech features.

2) WORD FEATURE

Word feature, that is, the vector representation of the fused full-text semantic information corresponding to each word. The processed text is subjected to IntentBERT to obtain word features. Adding the Embeddings on the 10th, 11th, and 12th layers of IntentBERT to obtain the final word feature e_n^w. The equation is as follows.

$$e_n^w = Embedding_{10} + Embedding_{11} + Embedding_{12} \tag{2}$$

where $Embedding_{10}$, $Embedding_{11}$ and $Embedding_{12}$ represent the output of the sequence of the 10th, 11th and 12th hidden layers of the Transformer, respectively. To further improve the effect of intention recognition, keyword features are introduced to mine keywords in the text and greater weight are given to them. w represents the word features.

3) KEYWORD FEATURE

Keyword features are the vector representation of the key part of the whole sentence. Finding the right keywords plays a vital role in intention recognition. Rake-NLTK identifies the key phrases in the text by analyzing the frequency of words in the text and the co-occurrence with other words. The keyword extraction process is shown in Fig.5.

If a sentence s of length n is entered, e_n^k denotes the vectorization matrix, the equation is as follows.

$$e_n^k = tag_1 + tag_2 + tag_3 + \ldots + tag_i \tag{3}$$

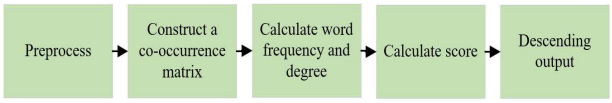

FIGURE 5. The keyword extraction process.

where tag_i is the keyword vector of the i-th word, $tag_i \in V^b$, and the dimension of the vector is b. k represents the keyword features.

After obtaining part-of-speech features, word features and keyword features respectively, the fusion vector is calculated. The equation is as follows.

$$e_n = concat[e_n^p, e_n^w, e_n^k] \tag{4}$$

where $concat$ represents the fusion of three features of part-of-speech e_n^p, word e_n^w, and keyword e_n^k.

C. ORTHOGONALITY CONSTRAINED MULTI-HEAD ATTENTION LAYER

Through the feature fusion layer, the vector after the fusion of part-of-speech features, word features and keyword features is obtained. The fusion vector is input into the orthogonality constrained multi-head attention layer, so that the model pays attention to the information of different sequences. The regularization term is used to reduce the position and representation redundancy between heads. And it guides the model to learn information that better fits the contextual features.

The orthogonality constrained multi-head attention network enables the model to jointly focus on the information at different locations. The attention mechanism calculates the attention distribution on the given information. And the weighted average of all input information according to the attention distribution is calculated.

Inputting text sequence $X = \{x_1, x_2, \ldots, x_n\}$, $X \in R^{n*d}$, the hidden state $H = \{h_1, h_2, \ldots, h_n\}$ is obtained by IntentBERT. Multi-head attention is used to learn context-related text features. H is used to obtain Q, K, V. And the contextually relevant text features A is calculated. The equation is as follows.

$$A = softmax(\frac{QK^T}{\sqrt{d_K}}), \quad A \in R^{T*d} \tag{5}$$

where Q, K, V denotes query, key and value respectively. Key uses query as the basis, and after calculation, the attention weight for each key is obtained. The final result is obtained by weighted summation of value.

The hidden layer vector H is learned by the initial IntentBERT. And the context-sensitive vector A is learned by the multi-head attention mechanism. Then the above two are combined to obtain the final discourse representation E. The equation is as follows.

$$E = H + A \tag{6}$$

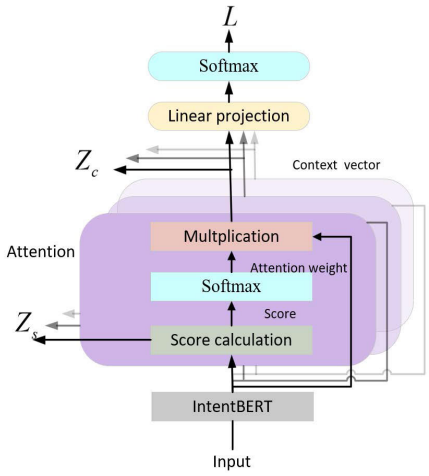

FIGURE 6. Orthogonality constrained multi-head attention frame.

where H is a vector representation of the hidden layer learned by the IntentBERT model. And A is a vector representation containing contextual information after the orthogonality constrained multi-head attention layer.

Given the source statement x and its intention y, the model is trained to maximize the probability of intention recognition. The orthogonal constraint is introduced to reduce the position and representation redundancy between the attention heads. And the diversity between the attention heads can be enhanced. $J(\theta)$ represents the loss function, the equation is as follows.

$$J(\theta) = argmin\{L(y|x;\theta) + \alpha * Z_c + \beta * Z_s\} \tag{7}$$

where x is the source statement, and y is the intention. a and β are hyperparameters, and the value are both set to 1.0. The orthogonal constraint Z_c and Z_s guide the relevant attention components to capture different features from the corresponding projection space. *argmin* returns the index of the minimum value. $L(.)$ works like L1 regularization and L2 regularization, but it does not introduce new parameters and does not affect the training of standard model parameters. Z_c represents the context vector regularization term, and Z_s represents the score vector regularization term.

The regularization term is introduced to make each head focus on different aspects of the sentence. And all words are covered by the head. The regularization term and the loss value are optimized together to further improve the performance of the model. The network parameters that minimize the cross-entropy are found by generating contextual outputs with minimal redundancy from each other. The orthogonality constrained multi-head attention frame diagram is shown in Fig.6.

Regularization means that the context vector and the score vector are orthogonal to each other. The orthogonality between the context vector and the score vector makes the attention heads have less redundancy. The equations of the context vector regularization term Z_c and score vector regularization term Z_s are as follows.

$$Z_c = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{H(H-1)}||C^{(n)T}C^{(n)} - I_H||_F^2 \tag{8}$$

$$Z_s = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{H(H-1)}||S^{(n)T}S^{(n)} - I_H||_F^2 \tag{9}$$

where n is the sample index, and H is the number of attention heads. $||.||_F^2$ represents the Frobenius norm. $C^{(n)}$ represents the context matrix, and $S^{(n)}$ represents the scoring matrix.

The equations of context matrix $C^{(n)}$ and score matrix $S^{(n)}$ are as follows.

$$C^{(n)} = [c_1^{(n)}, c_2^{(n)}, \ldots c_H^{(n)}] \tag{10}$$

$$S^{(n)} = [s_1^{(n)}, s_2^{(n)}, \ldots s_H^{(n)}] \tag{11}$$

$C^{(n)}$ and $S^{(n)}$ are composed of normalized context vector c_i and normalized score vector s_i, respectively.

D. FULLY CONNECTED LAYER

Intention recognition is essentially a classification task. The Sigmoid function is used to show the results of multi-classification in the form of probability. In the final output of the TMH-IntentBERT model, Sigmoid is selected as the classifier to predict the intention.

Specifically, given N different classes of labels, a linear layer is added as the classifier. $p(y|h_i)$ represents the probability of the intention label. The equation is as follows.

$$p(y|h_i) = sigmoid(Wh_i + b) \in R^N \tag{12}$$

where $h_i \in R^d$ is the characteristic representation of x_i given by [CLS] token. $W \in R^{N*d}$ and $b \in R^N$ are parameters of the linear layer. Model parameter $\theta = \{\varphi, W, b\}$, φ is a parameter of BERT, which is obtained by training the cross entropy loss function on $D_{source}^{labeled}$. $\theta*$ denotes the cross entropy loss function, and the equation is as follows.

$$\theta* = argminL_{ce}(D_{source}^{labeled};\theta) \tag{13}$$

where *argmin* returns the minimum index. $D_{source}^{labeled}$ represents labeled data and θ is the parameter of the model.

The Sigmoid function expression is as follows.

$$S(x) = \frac{1}{1+e^{-x}} \tag{14}$$

Through the sigmoid function, the output value of multi-classification can be converted into a probability distribution in the range of [0,1].

The pseudo-code of the TMH-IntentBERT Model is shown in Algorithm 1.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATASETS

In this paper, the model construction and experiment are carried out on the cloud server. Nvidia GeForce RTX3090 Ti GPU is used for training. BERTbase is used as encoder, and

Algorithm 1 TMH-IntentBERT Model

Input: intent dataset, learning rate lr, max_length max_len, dropout, epoch, batch-size.
Output: the probability of intent $p(y|h_i)$

```
text is obtained by stop words removal, word segmentation and disrupting data order
for e in range(1, epoch+1) do
    for i,batch in enumerate(dataloder,1) do
        X_pos ⇐ get_pos_features(text)
        X_word ⇐ get_word_features(text)
        X_kw ⇐ get_keyword_features(text)
        final_feature ⇐ X_pos + X_word + X_kw
        A = mha(final_features)
        E ⇐ final_features + A
        if valAcc >= valBestAcc then
            Accumulatestep+ = 1
            if accumulateStep > self.patience/2 then
                earlystop
            end if
        end if
        get best model MHA − IntentBERT
    end for
end for
x = MHA − IntentBERT(text)
p(y|h_i) = sigmoid(x)
return p(y|h_i)
```

TABLE 1. Dataset statistics.

Dataset	Intention field and quantity	Number
OOS	10 field 150 intent	23700
HWU64	21 field 64 intent	25716
BANGKING77	1 field 77 intent	13083
MCID	1 field 16 intent	1745
HINT3	3 field 51 intent	2011

Adam is used as optimizer. Python programming language and Pytorch framework are used for experiments.

Five datasets are used by the TMH-IntentBERT model. In order to train IntentBERT, BERT is pre-trained on HWU64 [37] and OOS [38]. Both datasets contain multiple domains and provide rich learning resources. The OOS dataset contains 150 intent categories in 10 domains. The HWU64 dataset contains 25716 examples of 64 intentions in 21 domains. Validation is performed on the remaining three datasets. Among them, BANKING77 [39] is a single domain intent detection dataset, including 13083 customer information such as complaints and problems sent to banks. The dataset contains 77 intents and focuses on fine-grained single-domain intent detection. The MCID dataset [40] contains 16 specific intentions and 4 languages for COVID-19 discourse intention detection tasks. HINT3 [41] covers 51 intent categories in three domains. The dataset statistics used in the experiment are shown in Table 1.

TABLE 2. Experimental hyperparameter value setting.

hyperparameter	parameter value
dropout	0.1
max_len	26
lr	1e-6
batch size	64
epoch	10

B. EVALUATION INDICATORS

The classification performance was evaluated by C-way and K-shot tasks. For each task, C classes are randomly selected, and K samples are extracted from each class to train the classifier. And then an additional 5 samples are extracted from each class as queries to evaluate. Take the average of 500 such tasks as accuracy.

In order to verify the effectiveness of the model classification, the *Accuracy*(hereinafter referred to as *Acc*), *precision*(hereinafter referred to as *Pre*), $F1$ value(hereinafter referred to as *Fsc*) and *AUROC* are used to evaluate the performance. *AUROC* is a metric used to evaluate the performance of a classification model. The *ROC* curve is a graphical representation of the rate of true positives versus the rate of false positives at different threshold settings. The *AUROC* is computed as the area under the *ROC* curve. The equation is as follows.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Pre = \frac{TP}{TP + FP} \tag{16}$$

$$Rec = \frac{TP}{TP + FN} \tag{17}$$

$$Fsc = \frac{2 * Pre * Rec}{Pre + Rec} \tag{18}$$

where TP represents the number of texts in which the actual label is true and the model intent recognition result is also true. FN represents the number of texts that the actual label is true, but the model intent recognition result is false. FP denotes the number of text that the actual label is false, but the model intent is recognized as true. TN indicates that the actual label is false, and the model intention recognition result is also the number of false texts.

C. EXPERIMENTAL PARAMETER SETTING

On the banking77 dataset (2 shot), the effects of dropout, max_length(hereinafter referred to as *max_len*), learning rate(hereinafter referred to as *lr*) and batch size on the *Acc*, *Pre*, *Fsc*, and *AUROC* of intention recognition are compared. The experimental hyperparameters are set as shown in Table 2.

1) THE EFFECT OF DROPOUT ON THE TMH-INTENTBERT MODEL

Dropout refers to temporarily discarding neural network units from the network according to a certain probability

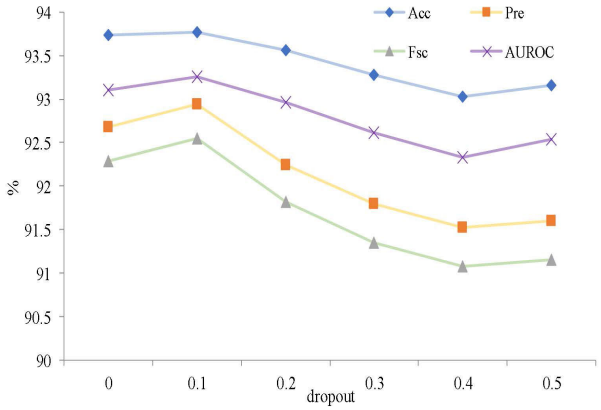

FIGURE 7. The effect of dropout on the TMH-IntentBERT model.

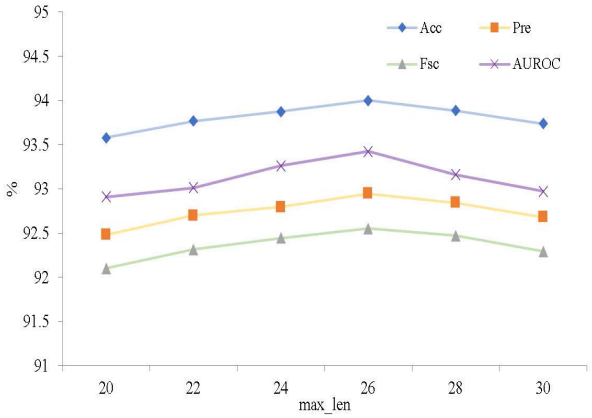

FIGURE 8. The effect of *max_len* on the TMH-IntentBERT model.

during the training of deep learning networks. Dropout can improve the over-fitting effect of the model and enhance the generalization ability of the model. On the banking77 dataset (2 shot), the impact of dropout on *Acc*, *Pre*, *Fsc*, and *AUROC* is shown in Fig.7.

It can be seen from Fig.7 that when dropout is 0.1, the *Acc*, *Pre*, *Fsc* and *AUROC* reach the maximum. As the dropout becomes larger, the effect becomes worse. At this time, the model appears overfitting. So the dropout of the TMH-IntentBERT model is 0.1.

2) THE EFFECT OF *MAX_LEN* ON THE TMH-INTENTBERT MODEL

The *max_len* is the maximum length, which is the parameter of part-of-speech features, keyword features and word segmenter. When the data is less than the *max_len*, the zero padding operation is performed. When the data length exceeds the *max_len*, the interception operation is performed. Therefore, too large or too small *max_len* will affect the performance of the model. On the banking77 dataset (2 shot), the effect of *max_len* on *Acc*, *Pre*, *Fsc*, and *AUROC* is shown in Fig.8.

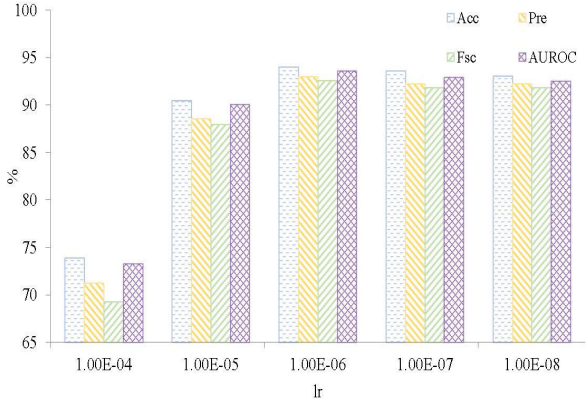

FIGURE 9. The effect of *lr* on TMH-IntentBERT model.

It can be seen from Fig.8 that when the *max_len* is 26, the effect is the best. This is because most of the length of the dataset is 26. If *max_len* is too small, it will be automatically truncated, resulting in some information missing. If *max_len* is too large, it will automatically zero and add useless information. Therefore, the *max_len* of the TMH-IntentBERT model is 26.

3) THE EFFECT OF *LR* ON THE TMH-INTENTBERT MODEL

lr directly affects the convergence state of the model. If the *lr* is too large, the model will not converge. If the *lr* is too small, the model will converge too slowly or cannot learn. The appropriate *lr* is the key to improving the effectiveness of the model. On the banking77 dataset (2 shot), the impact of the *lr* on *Acc*, *Pre*, *Fsc*, and *AUROC* is shown in Fig.9.

If the *lr* is too small, the convergence speed is very slow. It will not converge to the minimum value if the *lr* is too large. Therefore, selecting the appropriate *lr* is helpful to improve the training speed of the model. It can be seen from Fig.9 that the model works best when the *lr* is 1e-6. The *Acc* is 93.99%, the *Pre* is 92.94%, the *Fsc* is 92.55%, and the *AUROC* is 93.56%. Therefore, the *lr* of the TMH-IntentBERT model is 1e-6.

4) THE EFFECT OF BATCH SIZE ON THE TMH-INTENTBERT MODEL

The batch size affects the generalization performance of the model. To further improve the performance of the model, batch size is a very critical parameter. Within a certain range, increasing batch size will improve the stability of convergence and reduce training time, but the generalization ability of the model may be reduced. In order to verify the effect of batch size on the TMH-IntentBERT model, on the banking77 dataset (2 shot), the effect of batch size on *Acc*, *Pre*, *Fsc*, and *AUROC* is shown in Fig.10.

It can be seen from Fig.10 that as the batch size becomes larger, the *Acc*, *Pre*, *Fsc*, and *AUROC* of the model increase first and then decrease. When the batch size is 64, the *Acc*, *Pre*, *Fsc*, and *AUROC* of the model reach

TABLE 3. Selection of MHA-IntentBERT ablation experimental model.

Experiment number	Model	Feature fusion		
		POS features	KW features	OCM
Experiment1	POSIM	✓		
Experiment2	KWIM		✓	
Experiment3	POS+KWIM	✓	✓	
Experiment4	OCMHAIM			✓
Experiment5	POS+OCMHAIM	✓		✓
Experiment6	KW+OCMHAIM		✓	✓

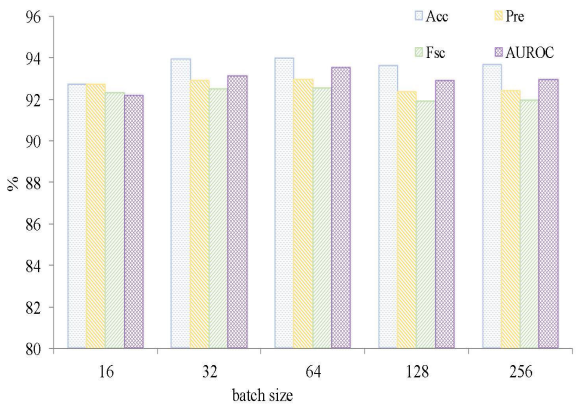

FIGURE 10. The effect of batch size on TMH-IntentBERT model.

the maximum, which are 93.99%, 92.94%, 92.55%, and 93.51% respectively. Therefore, when the batch size is 64, the TMH-IntentBERT model works best.

D. ABLATION EXPERIMENTS

Compared with the classical few-shot intent recognition method, a feature fusion module and an orthogonality constrained multi-headed attention module are added to the TMH-IntentBERT model. To verify the influence of each part of the TMH-IntentBERT model on the overall effect, the ablation experiments of MHA-IntentBERT on feature fusion and orthogonality constrained multi-head attention(OCM) are designed. Among them, feature fusion is mainly to fuse the part-of-speech features(POS features) and the keyword features(KW features) with word features. The ablation experimental models were selected as shown in Table 3.

In order to verify the effect of part-of-speech features and keyword features, the ablation experiment of feature fusion is carried out. The ablation experiment results are shown in Table 4.

As can be seen from Table 4 that the part-of-speech features achieve better results compared to keywords in terms of *Acc*, *Pre*, *Fsc*, and *AUROC*. This is due to the fact that part-of-speech features takes into account more fine-grained features not considered in the original model, which can effectively identify the semantics of polysemous words and improve the *Acc* of the model's intent recognition to a greater extent.

In order to verify the effect of orthogonality constrained multi-head attention, the ablation experiment of orthogonality constrained multi-head attention is carried out. The ablation experiment results are shown in Table 5.

It can be seen from Table 4 and Table 5 that compared with the multi-channel feature fusion module, the *Acc*, *Pre*, *Fsc*, and *AUROC* of the orthogonality constrained multi-head attention module are higher. In the absence of part-of-speech features and keyword features, the *Fsc* is reduced by 0.58%. The *Fsc* is reduced by 1.65% in the absence of the orthogonality constrained multi-headed attention module. In general, the two modules introduced in this paper have played a positive and effective role in the accuracy of few-shot intention recognition(Take 5 way 2 shot as an example).

The results of ablation experiments show that each additional feature can effectively improve the performance of the model based on the original text features. The more types of additional features, the deeper semantic information can be mined, and the better the model effect. The orthogonality constrained multi-head attention module has the greatest improvement in the *Acc* of the model. Because the multi-head attention mechanism can better represent the latent vector of the sentence and obtain the context information. The use of a multi-head attention mechanism can enable the model to learn more fully. And it effectively prevents the model from excessively focusing on its position when encoding the information at the current position. At the same time, regularization is used to reduce the position and representation redundancy between heads. And it enhance the diversity between attention heads. Based on this, the TMH-IntentBERT model achieves the best *Acc*, *Pre*, *Fsc*, and *AUROC* on the three datasets of banking77, mcid and hint3.

E. COMPARISON EXPERIMENTS

The TMH-IntentBERT model will be compared with the following four models for experiments.

1) CONVBERT [42]: The model dynamically generates convolution kernels using multiple input tokens. A span-based dynamic convolution operation is designed. By fine-tuning BERT on a corpus of nearly 700 million dialogues, the CONVBERT model is proposed.
2) TOD-BERT [43]: The model is improved based on BERT and applied to the dialogue domain. The NSP in

TABLE 4. Ablation experimental results of feature fusion.

Experiment number \ Evaluating indicator	Acc(%)	Pre(%)	Fsc(%)	AUROC(%)
Experiment1	91.26	92.56	90.81	90.78
Experiment2	91.19	92.50	90.08	90.32
Experiment3	91.45	92.77	91.03	90.10

TABLE 5. Ablation experimental results of orthogonality constrained multi-head attention.

Experiment number \ Evaluating indicator	Acc(%)	Pre(%)	Fsc(%)	AUROC(%)
Experiment4	92.53	93.56	92.10	91.95
Experiment5	92.65	93.69	92.21	92.18
Experiment6	92.55	93.59	92.15	92.10

TABLE 6. Results of comparative experiments(2 shot).

	Banking77				Mcid				hint3			
Model	Acc	Pre	Fsc	AUROC	Acc	Pre	Fsc	AUROC	Acc	Pre	Fsc	AUROC
IntentBERT	89.16	91.46	89.69	90.15	83.03	84.82	82.20	83.54	84.28	86.95	84.41	85.24
ConvBERT	69.71	72.37	68.20	70.63	70.26	72.72	68.85	70.45	70.62	73.16	69.13	72.05
TODBERT	78.06	80.47	76.77	78.95	68.85	70.39	67.23	68.96	75.65	78.06	74.40	77.26
WKR	54.23	56.50	52.32	55.54	46.37	46.87	44.01	46.02	48.23	50.15	46.16	48.68
DFT++	90.79	93.13	91.33	91.28	84.28	86.10	83.04	84.83	85.96	88.02	85.83	86.45
TMH-intentBERT	**92.68**	**93.74**	**92.29**	**92.32**	**85.07**	**86.68**	**84.24**	**85.76**	**86.99**	**88.68**	**86.22**	**87.59**

[1] The best results are highlighted.

TABLE 7. Results of comparative experiments(10 shot).

	Banking77				Mcid				hint3			
Model	Acc	Pre	Fsc	AUROC	Acc	Pre	Fsc	AUROC	Acc	Pre	Fsc	AUROC
IntentBERT	94.42	95.04	94.33	94.53	91.22	91.35	90.09	91.83	91.99	92.94	90.83	92.46
ConvBERT	82.40	84.39	81.85	83.56	80.96	82.66	80.40	81.56	82.64	84.77	82.08	83.64
TODBERT	88.34	90.01	87.97	88.95	82.02	83.81	81.45	82.86	87.25	89.20	86.88	88.15
WKR	71.43	73.89	70.59	72.54	59.48	61.25	58.15	60.34	64.68	67.29	63.68	65.57
DFT++	95.89	96.03	95.76	95.83	92.31	92.99	91.65	92.05	93.74	94.55	93.43	93.88
TMH-intentBERT	**96.66**	**97.21**	**96.58**	**96.93**	**93.04**	**94.05**	**92.86**	**93.46**	**94.37**	**95.28**	**94.22**	**95.02**

[1] The best results are highlighted.

BERT is replaced with RCL (response contrastive loss) while a dataset from the task-oriented dialogue domain is used for training.

3) WikiHowRoBERTa(WKR) [44]: The model further pre-trains RoBERTa in the WikiHow database, and a pre-trained intention recognition model is formed. In the end, a linear classification layer with cross-entropy loss is added to calculate the possibility of each intention.
4) IntentBERT [13]: About 1000 labeled data is used to fine-tune BERT, and randomly selects a part of the unlabeled discourse for joint pre-trained. The IntentBERT model is proposed for few-shot intent detection.
5) DFT++ [45]: The model proposes a context enhancement method that utilizes sequential self-distillation to improve the accuracy of the model. There is no excessive reliance on external databases and the overfitting problem is solved.

CONVBERT, TOD-BERT and WikiHowRoBERTa all require a lot of data and high computational costs. About 1000 labeled discourses from the public dataset are used by IntentBERT for standard supervised training. And IntentBERT is applied to the target domain that is significantly different from the pre-trained data for few-shot intent recognition. Based on IntentBERT, the TMH-IntentBERT algorithm is proposed, which adds multi-channel feature fusion and orthogonality constrained multi-head attention module to improve the accuracy of intention recognition.

Referring to the form of result presentation of intentBERT [13], the *Acc*, *Pre*, *Fsc*, and *AUROC* of the TMH-IntentBERT model compared with CONVBERT, TOD-BERT, WikiHowRoBERTa, IntentBERT and DFT++ on banking77, mcid and hint3 datasets in 2 shot and 10 shot are shown in Table 6 and Table 7.

From Table 6 and Table 7, it can be seen that the *Acc*, *Pre*, *Fsc*, and *AUROC* of the TMH-IntentBERT model show better performance on the three datasets of banking77, mcid

and hint3. Compared with DFT++, the model with the highest *Acc*, *Pre*, *Fsc*, and *AUROC* among CONVBERT, TOD-BERT, WikiHowRoBERTa, IntentBERT and DFT++, the TMH-IntentBERT model improves the *Acc* by 1.89%, 0.79% and 1.03% in the 2-shot case, and the *Fsc* by 0.96%, 1.20%, and 0.39%, respectively. In the case of 10 shot, the *Acc* increased by 0.77%, 0.73% and 0.63% respectively, and the *Fsc* increased by 0.82%, 1.21% and 0.79% respectively.

The mask language model is used by CONVBERT to fine-tune the unpacked BERT for 4 epochs. The results show that large-scale pre-trained on open domain dialogue data can be effectively transferred to task-oriented dialogue tasks. The language patterns between ordinary text and task-oriented conversations differ a lot. Aiming at the differences, the task-oriented conversation datasets are used to model the TOD-BERT. In the field of intention recognition, TOD-BERT is stronger than strong baselines such as BERT. The WikiHowRoBERTa model is trained on WikiHow and works well with very few samples in multiple languages. A small part of the marked discourse of the public dataset is used by IntentBERT to simply fine-tune BERT. And good results are produced in novel fields.

The part-of-speech features, word features and keyword features of the text are fused. The information contained in the data is fully mined. The fine-grained features of the text are fully considered by TMH-IntentBERT. To prevent the model from excessively focusing on its position when encoding the information at the current position, the multi-head attention mechanism is used. While more sufficient learning is gained by the model. Regularization is used between heads to reduce position and representation redundancy and further improve the accuracy of intent recognition.

V. CONCLUSION

With the coming of the era of intelligent dialogue, how to accurately identify the user's intention is a hot topic of current research. In this paper, the Triple Channel IntentBERT and Orthogonality Constrained Multi-Head Attention Model is proposed. Based on IntentBERT, multi-channel feature fusion technology is used to fuse the part-of-speech features, word features and keyword features of the dataset to make full use of the data. Feature learning is performed through the prior information tutoring model provided by the dataset. The orthogonality constrained multi-head attention mechanism is used to enhance the correlation of the output sequence features of the model. And it guides the model to learn a more contextual representation. The experimental results show that the TMH-IntentBERT model has a minimum increase of 0.63%, 0.73%, 0.79% and 1.10% in *Acc*, *Pre*, *Fsc* and *AUROC* compared with the baseline model CONVBERT, TOD-BERT, WikiHowRoBERTa, IntentBERT and DFT++, respectively. It indicates that the TMH-IntentBERT model has superior intention recognition accuracy.

The part-of-speech features, word features and keyword features are combined in the TMH-IntentBERT model. And the orthogonality constrained multi-head attention mechanism is used to make full use of the model hidden layer information. In the dialogue system, intent recognition and slot filling are two closely related tasks. In the future, how to jointly model intent recognition and slot filling will be the key consideration.

ACKNOWLEDGMENT

The authors look forward to the insightful comments and suggestions of the anonymous reviewers and editors, which will go a long way towards improving the quality of this paper.

DECLARATIONS

- Conflict of interest The authors declare that they have no conflict of interest.
- Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.
- Informed consent Informed consent was obtained from all individual participants included in the study.
- Authors' contributions Di Wu and Yuying Zheng wrote the main manuscript, Peng Cheng realised the experimental part. All authors reviewed the manuscript.

REFERENCES

[1] A. Yehudai, M. Vetzler, Y. Mass, K. Lazar, D. Cohen, and B. Carmeli, "QAID: Question answering inspired few-shot intent detection," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–13.

[2] A. Benayas, M. A. Sicilia, and M. Mora-Cantallops, "Automated creation of an intent model for conversational agents," *Appl. Artif. Intell.*, vol. 37, no. 1, Dec. 2023, Art. no. 2164401.

[3] R. Zhang, S. Luo, L. Pan, Y. Ma, and Z. Wu, "Strengthened multiple correlation for multi-label few-shot intent detection," *Neurocomputing*, vol. 523, pp. 191–198, Feb. 2023.

[4] W. A. Abro, G. Qi, M. Aamir, and Z. Ali, "Joint intent detection and slot filling using weighted finite state transducer and BERT," *Int. J. Speech Technol.*, vol. 52, no. 15, pp. 17356–17370, Dec. 2022.

[5] K. Zhao, X. Jin, and Y. Wang, "Survey on few-shot learning," *J. Softw.*, vol. 32, no. 2, pp. 349–369, 2020.

[6] J. Ramanand, K. Bhavsar, and N. Pedanekar, "Wishful thinking-finding suggestions and 'buy' wishes from product reviews," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 54–61.

[7] B. Thomson, *Statistical Methods for Spoken Dialogue Management*. Berlin, Germany: Springer, 2013.

[8] Z. Yang, L. Wang, and Y. Wang, "Application research of deep learning algorithm in question intention classification," *Comput. Eng. Appl.*, vol. 55, no. 10, pp. 154–160, 2019.

[9] V. Hudecek and O. Dusek, "Are large language models all you need for task-oriented dialogue?" in *Proc. 24th Meeting Special Interest Group Discourse Dialogue*, 2023, pp. 216–228.

[10] X. Zheng and J. ren, "Intention recognition and classification based on bert-FNN," *Computer Modernization*, no. 7, pp. 71–76, 2021.

[11] J. Kapociute-Dzikiene, A. Salimbajevs, and R. Skadins, "Monolingual and cross-lingual intent detection without training data in target languages," *Electronics*, vol. 10, no. 12, p. 1412, Jun. 2021.

[12] R. Chen, H. Li, G. Yan, Z. Wang, and H. Peng, "Target intent recognition method based on evidence fusion in TimeSeries networks," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Oct. 2022, pp. 1–6.

[13] H. Zhang, Y. Zhang, L. M. Zhan, J. Chen, G. Shi, X. M. Wu, and A. Y. Lam, "Effectiveness of pre-training for few-shot intent classification," in *Proc. Findings Assoc. Comput. Linguistics, Findings ACL*, 2021, pp. 1114–1120.

[14] D. Chen, "Triple-channel feature mixed sentiment analysis based on attention," *Comput. Eng. Des.*, vol. 43, nos. 546–552, Jan. 2022.

[15] L. Qiu, Y. Chen, H. Jia, and Z. Zhang, "Query intent recognition based on multi-class features," *IEEE Access*, vol. 6, pp. 52195–52204, 2018.
[16] X. Z. Y. Hua and J. Liu, "Few-shot object detection based on feature fusion," *Comput. Sci.*, vol. 50, no. 2, pp. 209–213, 2023.
[17] X. Zhou, T. Zhang, C. Cheng, and S. Song, "Dynamic multichannel fusion mechanism based on a graph attention network and BERT for aspect-based sentiment classification," *Int. J. Speech Technol.*, vol. 53, no. 6, pp. 6800–6813, Mar. 2023.
[18] L. Yu, Y. Tian-Tian, Z. Dong-Sheng, and W. Xiao-Peng, "Natural scene text detection based on attention mechanism and deep multi-scale feature fusion," *J. Graph.*, vol. 44, no. 3, p. 473, 2023.
[19] C. Xia, C. Zhang, X. Yan, Y. Chang, and P. S. Yu, "Zero-shot user intent detection via capsule neural networks," 2018, *arXiv:1809.00385*.
[20] S. Xue and F. Ren, "Intent-enhanced attentive BERT capsule network for zero-shot intention detection," *Neurocomputing*, vol. 458, pp. 1–13, Oct. 2021.
[21] C. Liu and X. Xu, "AMFF: A new attention-based multi-feature fusion method for intention recognition," *Knowl.-Based Syst.*, vol. 233, Dec. 2021, Art. no. 107525.
[22] L. Wang, H. Yang, F. Li, W. Yang, and Z. Zou, "Intent detection model based on dual-channel feature fusion," in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, vol. 6, Mar. 2022, pp. 1862–1867.
[23] M. Henderson, I. Casanueva, N. Mrksic, P.-H. Su, T.-H. Wen, and I. Vulic, "ConveRT: Efficient and accurate conversational representations from transformers," 2019, *arXiv:1911.03688*.
[24] J. Zhang, K. Hashimoto, W. Liu, C. Wu, Y. Wan, P. Yu, R. Socher, and C. Xiong, "Discriminative nearest neighbor few-shot intent detection by transferring natural language inference," in *Proc. EMNLP*, 2020, pp. 1–19.
[25] M. Hu, J. Peng, W. Zhang, J. Hu, L. Qi, and H. Zhang, "An intent recognition model supporting the spoken expression mixed with Chinese and English," *J. Intell. Fuzzy Syst.*, vol. 40, no. 5, pp. 10261–10272, Apr. 2021.
[26] B. Xu, Y. Ma, and H. Lin, "A hybrid deep neural network model for query intent classification," *J. Intell. Fuzzy Syst.*, vol. 36, no. 6, pp. 6413–6423, Jun. 2019.
[27] H. Nguyen, C. Zhang, C. Xia, and P. Yu, "Dynamic semantic matching and aggregation network for few-shot intent detection," in *Proc. EMNLP*, 2020, pp. 1–10.
[28] X. Wan, W. Zhang, M. Huang, S. Feng, and Y. Wu, "A unified approach to nested and non-nested slots for spoken language understanding," *Electronics*, vol. 12, no. 7, p. 1748, Apr. 2023.
[29] Z. Li, Q. Zhou, C. Li, K. Xu, and Y. Cao, "Improving BERT with syntax-aware local attention," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 645–653.
[30] Y. Ma, Z. Huang, J. Su, H. Shi, D. Wang, S. Jia, and W. Li, "A multi-channel feature fusion CNN-Bi-LSTM epilepsy EEG classification and prediction model based on attention mechanism," *IEEE Access*, vol. 11, pp. 62855–62864, 2023.
[31] J. Li, Z. Tu, B. Yang, M. R. Lyu, and T. Zhang, "Multi-head attention with disagreement regularization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2897–2903.
[32] M. Lee, J. Lee, H. J. Jang, B. Kim, W. Chang, and K. Hwang, "Orthogonality constrained multi-head attention for keyword spotting," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 86–92.
[33] Z. Hou, K. Ma, Y. Wang, J. Yu, K. Ji, Z. Chen, and A. Abraham, "Attention-based learning of self-media data for marketing intention detection," *Eng. Appl. Artif. Intell.*, vol. 98, Feb. 2021, Art. no. 104118.
[34] X. Hao, L. Wang, H. Zhu, and X. Guo, "Joint agricultural intent detection and slot filling based on enhanced heterogeneous attention mechanism," *Comput. Electron. Agricult.*, vol. 207, Apr. 2023, Art. no. 107756.
[35] P. Wei, B. Zeng, and W. Liao, "Joint intent detection and slot filling with wheel-graph attention networks," *J. Intell. Fuzzy Syst.*, vol. 42, no. 3, pp. 2409–2420, Feb. 2022.
[36] X. Huang, T. Ma, L. Jia, Y. Zhang, H. Rong, and N. Alnabhan, "An effective multimodal representation and fusion method for multimodal intent recognition," *Neurocomputing*, vol. 548, Sep. 2023, Art. no. 126373.
[37] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking natural language understanding services for building conversational agents," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*. Berlin, Germany: Springer, 2021, pp. 165–183.
[38] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, and L. Tang, "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 1311–1316.
[39] I. Casanueva, T. Temč inas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," in *Proc. 2nd Workshop Natural Lang. Process. Conversational AI*, 2020, pp. 38–45.
[40] A. Arora, A. Shrivastava, M. Mohit, L. S.-M. Lecanda, and A. Aly, "Cross-lingual transfer learning for intent detection of COVID-19 utterances," 2020, *arXiv:2006.03202*.
[41] G. Arora, C. Jain, M. Chaturvedi, and K. Modi, "HINT3: Raising the bar for intent detection in the wild," in *Proc. 1st Workshop Insights Negative Results NLP*, 2020, pp. 100–105.
[42] S. Mehri, M. Eric, and D. Hakkani-Tur, "DialoGLUE: A natural language understanding benchmark for task-oriented dialogue," 2020, *arXiv:2009.13570*.
[43] C.-S. Wu, S. C. H. Hoi, R. Socher, and C. Xiong, "TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 917–929.
[44] L. Zhang, Q. Lyu, and C. Callison-Burch, "Intent detection with wikihow," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 328–333.
[45] H. Zhang, H. Liang, L. Zhan, X.-M. Wu, and A. Y. S. Lam, "Revisit few-shot intent classification with PLMs: Direct fine-tuning vs. continual pre-training," 2023, *arXiv:2306.05278*.

DI WU received the Ph.D. degree in computer application technology from the School of Information Science and Engineering, Yanshan University. She is currently an Associate Professor with the School of Information and Electrical Engineering, Hebei University of Engineering. Her main research interests include data mining, natural language processing, software security analysis, and information retrieval.

YUYING ZHENG was born in Hebei, China, in 1999. She received the bachelor's degree from Hebei Normal University, in 2021. She is currently pursuing the master's degree with the School of Information and Electrical Engineering, Hebei University of Engineering. Her research interests include deep learning and NLP.

PENG CHENG was born in Chongqing, China, in 1997. He received the bachelor's degree from Chongqing Technology and Business University, in 2020. He is currently pursuing the master's degree with the School of Information and Electrical Engineering, Hebei University of Engineering. His research interests include deep learning and NLP.