aerospace

Article

Ernie-Gram BiGRU Attention: An Improved Multi-Intention Recognition Model for Air Traffic Control

Weijun Pan, Peiyuan Jiang, Zhuang Wang *, Yukun Li and Zhenlong Liao

College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China; wjpan@cafuc.edu.cn (W.P.); darcy@cafuc.edu.cn (P.J.); liyukun@cafuc.edu.cn (Y.L.); longmaoplus@cafuc.edu.cn (Z.L.)

* Correspondence: wangzhuang@cafuc.edu.cn

Abstract: In recent years, the emergence of large-scale pre-trained language models has made transfer learning possible in natural language processing, which overturns the traditional model architecture based on recurrent neural networks (RNN). In this study, we constructed a multi-intention recognition model, Ernie-Gram_Bidirectional Gate Recurrent Unit (BiGRU)_Attention (EBA), for air traffic control (ATC). Firstly, the Ernie-Gram pre-training model is used as the bottom layer of the overall architecture to implement the encoding of text information. The BiGRU module that follows is used for further feature extraction of the encoded information. Secondly, as keyword information is very important in Chinese radiotelephony communications, the attention layer after the BiGRU module is added to realize the extraction of keyword information. Finally, two fully connected layers (FC) are used for feature vector fusion and outputting intention classification vector, respectively. We experimentally compare the effects of two different tokenizer tools, the BERT tokenizer tool and Jieba tokenizer tool, on the final performance of the Bert model. The experimental results reveal that although the Jieba tokenizer tool has considered word information, the effect of the Jieba tokenizer tool is not as good as that of the BERT tokenizer tool. The final model's accuracy is 98.2% in the intention recognition dataset of the ATC instructions, which is 2.7% higher than the Bert benchmark model and 0.7–3.1% higher than other improved models based on BERT.

Keywords: transfer learning; multi-intention recognition; Ernie-Gram_BiGRU_Attention; air traffic control; tokenizer

check for updates

Citation: Pan, W.; Jiang, P.; Wang, Z.; Li, Y.; Liao, Z. Ernie-Gram BiGRU Attention: An Improved Multi-Intention Recognition Model for Air Traffic Control. *Aerospace* **2023**, *10*, 349. https://doi.org/10.3390/aerospace10040349

Academic Editor: Álvaro Rodríguez-Sanz

Received: 16 February 2023
Revised: 27 March 2023
Accepted: 29 March 2023
Published: 3 April 2023

1. Introduction

The International Civil Aviation Organization (ICAO) states in Doc 4444 PANS-ATM that the primary purpose of ATC is to prevent collisions between aircraft, prevent collisions between aircraft in the maneuvering area and obstructions in that area, and expedite and maintain an orderly flow of air traffic. To achieve this objective, ATC instructions are issued to aircraft in accordance with the provisions of the Air Traffic Services (ATS) Plan and other related documents. Therefore, ensuring the effective transmission of ATC instruction information is the key to air traffic safety. Currently, the Civil Aviation Administration of China (CAAC) is striving to promote the construction of intelligent ATC, which includes some key technologies, such as multimodal fusion, automatic ATC speech recognition, radiotelephony communication intention recognition, namely ATC instruction intention recognition, automatic response of ATC instructions, semantic verification of ATC instructions, etc. [1]. ATC instruction intention recognition refers to the deep learning technology that enables the computer to judge the intention contained in the ATC instructions. The computer identifies the intention of the ATC instructions and the intention repeated by the pilot to determine if there is a miscommunication. Intention recognition technology can also be applied to human–computer dialogue systems to improve the accuracy of response generation [2,3]. Through this technology, the computer in the air traffic controller (ATCO) training simulator can improve the accuracy of generating response instructions so as to better execute the corresponding actions in the instructions.

Intention recognition is defined as a classification problem, and the main classification data are the current input dialogue information [4]. In the early period, various scholars mainly used the method based on statistical learning. Haffner et al. proposed the support vector machine (SVM) model, which has achieved good results in the field of classification and has been widely applied to intention recognition tasks [5]. Hakkani-Tur et al. used heterogeneous features extracted from semantic and syntactic graphs of user utterances to carry out intention recognition [6]. Kim improved the performance of the model in intention recognition tasks by improving the word embedding method [7]. Jeong et al. proposed a triangular conditional random field [8], which carries out intention recognition by adding an additional random variable to the standard conditional random field [9]. However, the features of the above methods are determined manually based on experience, which means the methods based on statistical learning have some problems such as heavy dependence on the size of the datasets, sparse extracted feature vector, and the extracted feature vector cannot effectively represent the semantic information of a short text.

The intention recognition method based on deep learning can solve the above problems better. For the first time, Kim proposed applying the convolutional neural networks (CNN) model for visual tasks to text classification tasks, built a convolutional neural networks for sentence classification (TextCNN) model, and conducted a series of convolution experiments based on Word2Vec architecture. The results show that simple CNN with one layer of convolution can perform better [10]. However, TextCNN has a problem in terms of selecting an appropriate convolution window. If the convolution window selected is too large, the computational complexity will become higher due to the large number of parameters; if the convolution window selected is too small, it will cause the loss of semantic information to some extent. In order to overcome this shortcoming of TextCNN, Wang et al. proposed recurrent convolutional neural networks (RCNN), which can overcome information loss and maximize the extraction of context information. Compared with TextCNN, it achieves a better effect [11]. Zhou et al. built a travel consumption intention model based on CNN and long short-term memory (LSTM), which makes up for the shortcoming of CNN's inability to extract deep meaning and semantic information [12]. Liu et al. used an attention mechanism to obtain important word information in sentences and combined it with a bidirectional circulatory neural network to extract sentence features to improve the accuracy of intention recognition [13]. However, the above intention recognition-based deep learning studies only consider single intention recognition and do not consider the case of multiple intentions in text sentences. For multi-intention recognition tasks, Lin et al. proposed a bidirectional long short-term memory (BiLSTM) model based on the self-attention mechanism for intention classification, in which the self-attention mechanism is used to obtain various semantic information of sentences [14]. With the proposal of BERT, this opens up a new way of thinking about natural language processing (NLP) tasks. Sun et al. apply BERT to text classification to study different methods based on BERT fine-tuning in text classification tasks, and the results show that BERT has excellent performance and huge potential in text classification [15]. Although the transfer learning model based on BERT has achieved amazing results in the task of natural language understanding (NLU), when the NLU task was carried out in the field of ATC, semantic information was still lost due to reasons stemming from the model itself, such as the model of BERT's failure to consider Chinese word information when training masked language model (MLM) tasks and the fact that the feature dimension of model output is limited, etc.

For the ATC instruction multi-intention recognition, keyword information extraction has a huge impact on the results of recognition, and choosing an appropriate word embedding method is also important. Therefore, the model we construct should effectively encode text information and capture keyword information in ATC instructions to improve the performance of the model. Using the Ernie-Gram model is a good way to overcome the limitations of BERT's pre-training model in word information extraction. The BiGRU module can capture global semantic information, and the attention mechanism can further extract instruction keyword information. However, it is not ideal to simply concatenate the results of each module. We use an FC to fuse the text vector output from the BiGRU module with the output from the attention layer, and splice the obtained fused vector with the classification word vector output by the Ernie-Gram model to obtain the final feature vector for multi-intention recognition. The improved model can extract semantic information from multiple levels, and expand the dimension of final classification word vector, significantly improving the accuracy of multi-intention recognition. The remainder of this paper is as follows. Section 2 highlights the difficulties in using multi-intention recognition techniques in this field and expounds the efforts of other researchers in the application of techniques in ATC. Section 3 describes the motivation of our model and the strategies adopted. In addition, we describe the principle and structure of each module in the model, and the overall framework of the model is presented. Section 4 introduces the characteristics of ATC instruction and the multi-intention recognition dataset of ATC instructions, compares the performance of different models in the target domain, and analyzes the experimental results. Section 5 summarizes the major findings and provides an outlook on our future work.

2. Challenges

In the past few years, many experts and scholars have conducted research on intention recognition, but most of their research focuses on the intention recognition of aircraft. Based on the constructed intention recognition model, they can predict the flight path of aircraft [16] or conduct situation awareness. In fact, the intention information contained in ATC instructions is very important for the safe operation of aircraft, so it is necessary to study the intention of ATC instructions in depth. At present, there are the following challenges in applying intention recognition technology to air traffic management:

(1) Model performance problem: Most of the intention recognition models currently studied are mainly used as a sub-module in the ATC monitoring system [17], so most scholars' work is mainly focused on improving the overall performance of the system. This results in a lack of in-depth research on the intention recognition sub-module. It is necessary to apply the latest technology to the field of ATC and build a model with high performance and good compatibility.

(2) Multiple intention recognition problem: At present, intention recognition based on deep learning technology is mainly aimed at single intention recognition. Since ATC instructions are characterized by short sentences, concise content, and no ambiguity, multi-intention recognition is a difficulty in this field. For example, the sentence shun feng/wu yao san guai/, lei da/kan dao/, shang sheng/dao/guai liang/bao chi only contains seven words and one word, but contains two intentions, which are aircraft identification and positioning and aircraft altitude adjustment. How to improve the accuracy of multi-intention recognition in short texts is a difficult problem that few people have studied.

(3) Fewer data and higher labeling costs: With the increasing number of deep learning model parameters, for supervised learning, a multi-intention recognition model with superior performance needs to be trained with a large amount of data. In the field of ATC, data acquisition is difficult due to the confidentiality of data. In addition, the acquired original ATC speech data can only be used after being marked by professionals, which brings great challenges to the application and development of deep learning technology in this field [18].

3. Methodology

3.1. Response Strategies to Challenges

For the first challenge, we conducted in-depth research on the ATC instruction intent recognition model. We studied mainstream intent recognition methods in other fields and, based on the characteristics of ATC instructions, constructed a high-performance ATC instruction multi-intention recognition model specifically for the air traffic control domain using advanced technical methods. For the second challenge, we employed a feature fusion strategy to construct an intention recognition model that can effectively extract local and global semantic information from short texts. At the base layer, we utilized a pre-trained model to encode text and extract global semantic features, while adopting bidirectional temporal neural networks and attention mechanisms to extract local semantic features. We compared and studied different feature fusion strategies, and selected the most suitable one to fuse global and local semantic features. The model we constructed can effectively handle ATC instruction texts containing multiple intents. For the third challenge, we adopted transfer learning to avoid the issue of model performance degradation caused by insufficient data. We improved the language model obtained by jointly pre-training on multiple tasks, and fine-tuned the improved model in the ATC domain, effectively alleviating the problem of poor performance in the small sample domain. On the basis of handling these challenges, we also performed an additional task. We validated the impact of replacing the character-based tokenizer with a word-based tokenizer on the performance of the pre-trained model, which lays the foundation for subsequent related research.

3.2. Model Introduction

3.2.1. Model Structure

The final model structure diagram we built is shown in Figure 1. The first layer of the model structure is the Ernie-Gram pre-training language model, which is responsible for encoding the input sentence. The BiGRU module is used for the global feature extraction of encoded sentence vectors. The Attention module selects the local information to focus on according to the extracted global information. FC is responsible for the feature fusion of extracted global information and local information. Finally, the fused feature vector and the CLS word vector output by Ernie-Gram are spliced together and fed into FC to realize the multi-intention classification of ATC instructions.

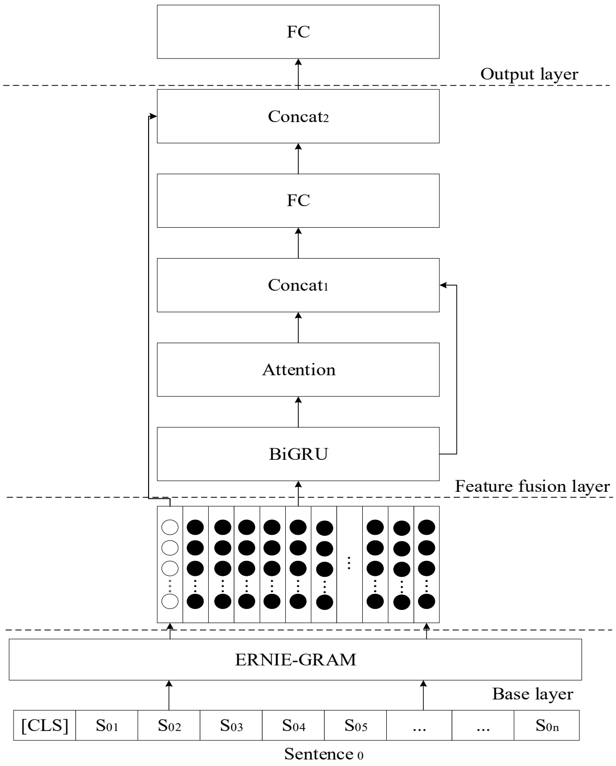


Figure 1. EBA model structure.

Supposing the input sentence sequence is S_0 = {s_{01}, s_{02}, ... , s_{0m}}, the calculation process of our model is as follows:

$$T, CLS = Ernie - Gram(S_0) \tag{1}$$

The *Ernie*-*Gram*() module encodes text information, T is the embedded representation of sentence S_0, and CLS represents the hidden state vector output by *Ernie*-*Gram* that contains the whole sentence information.

$$S_0', (\overrightarrow{H}, \overleftarrow{H}) = BiGRU(T) \tag{2}$$

S_0' is the further encoding of T by *BiGRU*() and $(\overrightarrow{H}, \overleftarrow{H})$ is the forward and backward hidden state vector of the output of the last layer of the *BiGRU* module.

$$H = concat(\overrightarrow{H}[-1], \overleftarrow{H}[1]) \tag{3}$$

The *concat*() function can concatenate vectors in a certain dimension. $\overrightarrow{H}[-1]$ and $\overleftarrow{H}[1]$ represent the forward final hidden state vector and the backward final hidden state vector, respectively.

$$K = Attention((H, S_0')) \tag{4}$$

$$G = Linear((K, H)) \tag{5}$$

The *Attention*() function and the *Linear*() function are used to compute the attention vector and the feature fusion of the vector, respectively. G represents the vector obtained by fusing the attention vector K with the final hidden state vector H.

$$I = concat(CLS, G) \tag{6}$$

I is the vector that is ultimately used for classification and is a combination of vector G and vector CLS.

$$y_{pre} = softmax(IW + b) \tag{7}$$

The *softmax*() function is used to calculate the probability of each intention. W and b represent the weight and bias of FC, respectively, and y_{pre} is the final output and is a probability vector.

The details of the architecture are shown in Table 1:

Table 1. Details of architecture.

Structural Order	Input Size	Output Size	Parameter Setup
Ernie-Gram	(32, 50)	(32, 50, 768), (32, 768)	Batch_size = 32, Max_length = 50
BiGRU	(32, 50, 768)	(32, 50, 768), (4, 32, 384)	Hidden_size = 384, Number_layers = 2
Extraction	(4, 32, 384)	(32, 768)	$\overrightarrow{H}[index = -1]$, $\overleftarrow{H}[index = 1]$
Attention layer	(32, 768)	(32, 768)	Extraction (32, 768) for q; BiGRU_output (32, 50, 768) for k and v;
Concat$_1$	(32, 768), (32, 768)	(32, 1536)	dimension = −1
FC	(32, 1536)	(32, 768)	FC (Concat$_1$)
Concat$_2$	(32, 768), (32, 768)	(32, 1536)	dimension = −1
FC	(32, 1536)	(32, 15)	Num_class = 15

3.2.2. Ernie-Gram Module

Ernie-Gram is a Chinese pre-training model improved by Baidu based on the BERT model. The input of this model is a single sentence or sentence pair, with additional segment embeddings and position embeddings that the model can learn by itself. The input structure diagram of the model is shown in Figure 2:

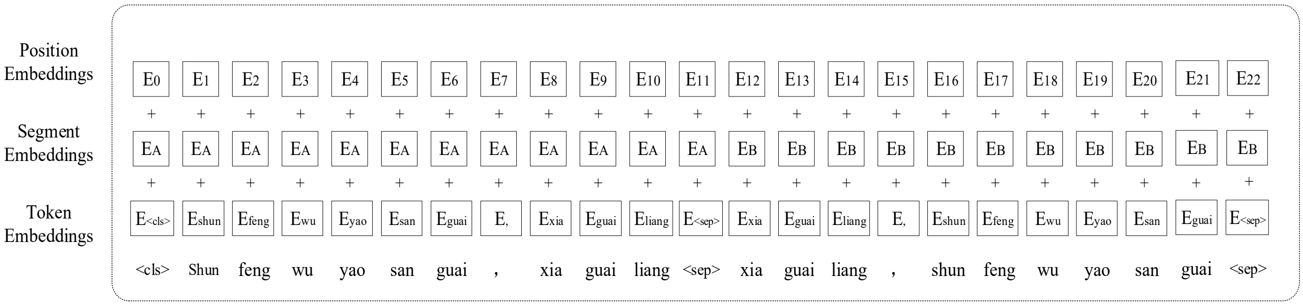
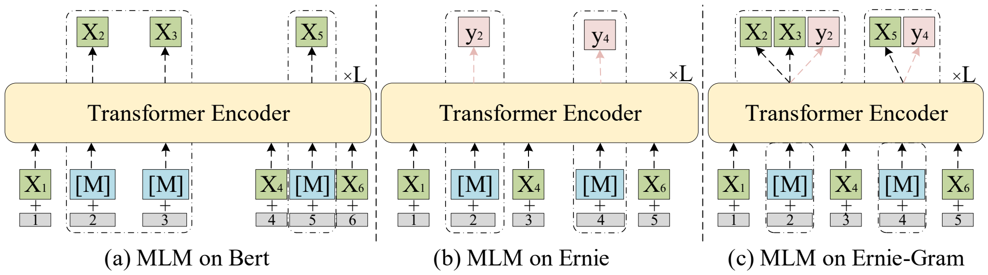

Position Embeddings
Segment Embeddings
Token Embeddings
E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16 E17 E18 E19 E20 E21 E22
EA EB
E<cls> Eshun Efeng Ewu Eyao Esan Eguai E, Exia Eguai Eliang E<sep> Exia Eguai Eliang E, Eshun Efeng Ewu Eyao Esan Eguai E<sep>
<cls> Shun feng wu yao san guai , xia guai liang <sep> xia guai liang , shun feng wu yao san guai <sep>

Figure 2. Input structure diagram of Ernie-Gram model.

The <cls> character is the classifier marker that starts the sentence. The <sep> character is a separator that splits two sentences and ends the sentence pair. E_A and E_B are segment embeddings of two sentences, respectively, and $E_{0–22}$ are position embeddings that are self-learnable.

3.2.3. The Difference between Ernie-Gram and BERT

When carrying out masked language model (MLM) pre-training tasks, the BERT model considers the representation of fine-grained text units [19], which, in Chinese, only means considering the concealment of a single word and predicting it. This processing method ignores a structured information of Chinese language, which is the word. For example, the masked character ‘qu’ is predicted based on the local co-occurrence of the three characters ‘xia’, ’men’, and ‘diao’. In this process, the model does not learn larger semantic units, such as the words, ‘Xiamen’ and ‘qu diao’, etc. The MLM strategy of the BERT model is shown in Figure 3a. The model of Ernie introduces knowledge enhancement learning based on the BERT model, and realizes better pretraining learning by masking successive N-Grams. Based on the Ernie model, Ernie-Gram takes into account the dependencies between N-Grams and uses an integrated N-Gram prediction mechanism. Ernie-Gram predicts masked N-Grams in both coarse-grained and fine-grained ways through carefully designed attention-masking measures, which enables the model to have stronger semantic modeling capabilities [20]. Figure 3b shows the MLM task strategy considering N-grams, and Figure 3c shows the MLM task strategy considering both N-grams and the direct dependencies within N-grams.



(a) MLM on Bert (b) MLM on Ernie (c) MLM on Ernie-Gram

Figure 3. Comparison of MLM task structures with different strategies.

3.2.4. BiGRU Module

BiGRU is a neural network model composed of forward and reverse GRUs. The GRU consists of two control units: reset gate unit R_t and update gate unit Z_t. The reset gate is used to remember or forget the hidden state information before the current moment, and the update gate is used to determine whether to update the hidden state of the current moment or use the hidden state of the previous moment as the hidden state of the current moment [21,22]. The GRU structure diagram is shown in Figure 4.

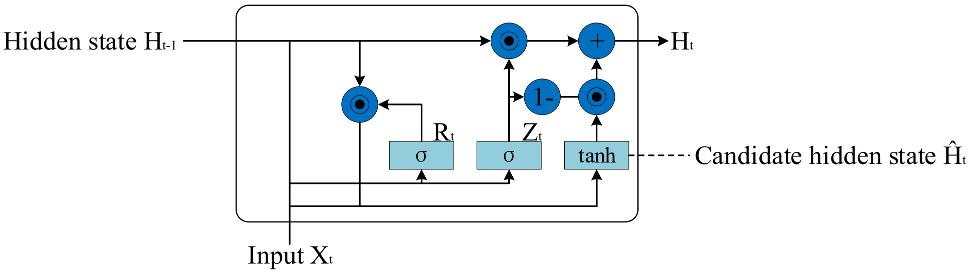


Figure 4. Comparison of MLM task structures with different strategies.

The specific calculation formula is as follows:

$$R_{\mathrm{t}} = \sigma(X_t W_{xr} + H_{t-1} W_{hr} + b_r) \tag{8}$$

$$Z_{\mathrm{t}} = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \tag{9}$$

$$\hat{H}_{\mathrm{t}} = \tanh(X_t W_{xh} + (R_t \odot H_{t-1}) W_{hh} + b_h) \tag{10}$$

$$H_{\mathrm{t}} = Z_{\mathrm{t}} \odot H_{t-1} + (1 - Z_{\mathrm{t}}) \odot \hat{H}_{\mathrm{t}} \tag{11}$$

R_t represents the reset gate, X_t represents the input at the current moment, W_{xr} represents the weight matrix of the reset gate at the current moment, H_{t-1} is the hidden state at the previous moment, W_{hr} is the reset gate's weight matrix of the hidden state before the current moment, and b_r is the bias of the reset gate. Z_t represents the update gate, $\odot$ represents the multiplication, σ represents the sigmoid activation function, and $\hat{H}_t$ represents the candidate hidden state.

GRU only considers the previous information, not the later information, when extracting the time series information. For the task whose input text is a complete statement, the GRU model will result in semantic information loss. The output of each step of BiGRU takes into account the combination of the forward-propagated hidden state of the previous step and the backpropagated hidden state of the later step of the current state, which enables the model to consider the semantic information of the entire sentence when making the output. The structure diagram of the BiGRU module is shown in Figure 5.

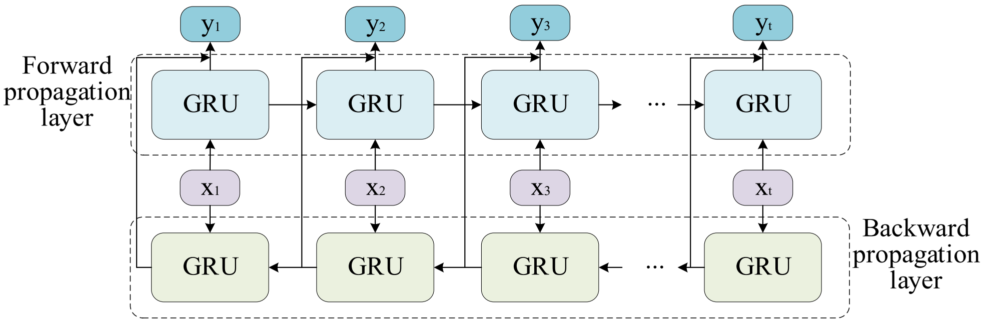


Figure 5. BiGRU structure diagram.

The formula of the BiGRU module is as follows:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \tag{12}$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}) \tag{13}$$

$$h_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \tag{14}$$

The *GRU*() function represents the nonlinear transformation of the input word vectors, encoding the word vectors into the corresponding *GRU* hidden state, w_t and v_t represent the weights of the forward and reverse hidden states corresponding to BiGRU at time t, respectively, and b_t represents the bias corresponding to the hidden state at time t.

3.2.5. Attention Mechanism

The purpose of the attention mechanism is to focus on the details according to the target. It overcomes the problem of information loss caused by too long a time series, and can be used to link the encoded hidden state vectors with the input encoded vectors, playing a role of highlighting the text keyword information. The core of the attention mechanism is the query vector q, the key vector k, and the value vector v. We can calculate the attention vector through the operation of the three vectors above [23,24]. Its principle is shown in Figure 6.

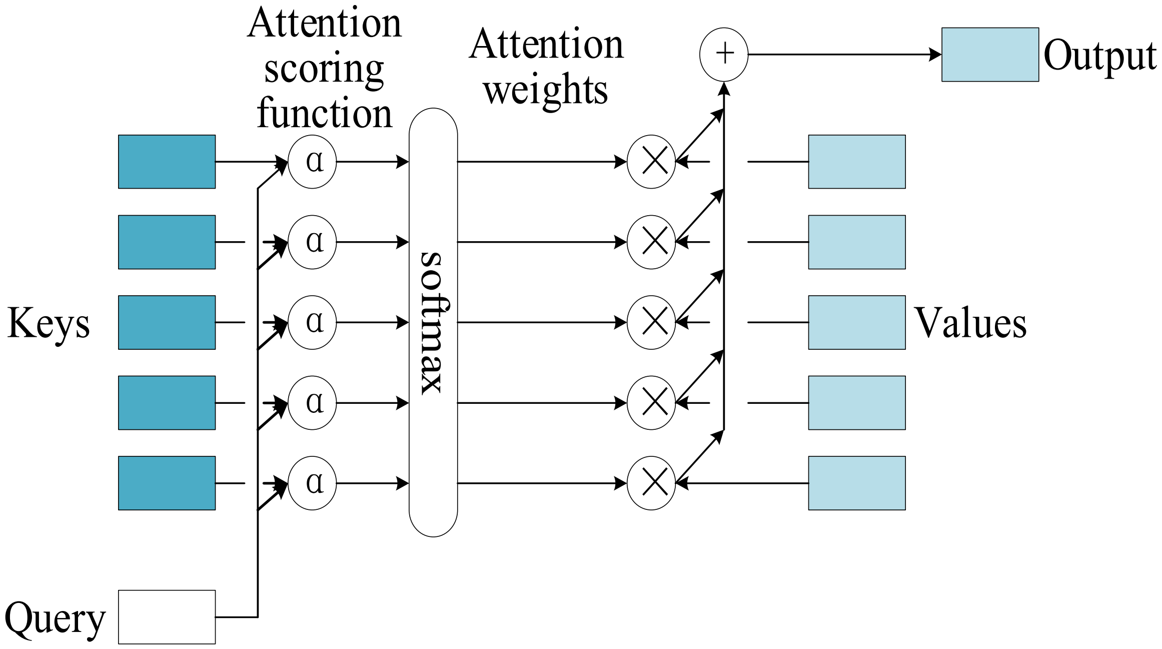


Figure 6. Schematic diagram of the attention mechanism.

The calculation process is as follows:

$$\mathrm{s}(k,q) = k^T q \tag{15}$$

$$\alpha(k_i,q) = \frac{e^{s(k_i,q)}}{\sum_{j=1}^{m} e^{s(k_j,q)}} \tag{16}$$

$$att(X,q) = \sum_{i=1}^{m} \alpha(k_i,q) v_i \tag{17}$$

In the three formulas above, q is the query vector with $q \in R^q$ and k is a time step in the input time series with $k \in R^q$. Key vector k is equal to the value vector v. The function $S()$ is used to calculate the attention score between the query vector and the key vector. $\alpha()$ is the attention distribution function. $att()$ is the expectation function, which is the expectation of the input vectors under the attention distribution.

3.3. Motivation and Details for Model Building

We found that, although referring to the current mainstream methods, only using the Ernie-Gram pre-training model to fine-tune can achieve the classification of multi-intention text [25,26]. However, we found that when we applied the above fine-tuning strategy to the ATC instruction text, the model's performance was not satisfactory. This is because, to a large extent, the text of the ATC instructions is a kind of natural language with particularity. Due to its unique professionalism and conciseness, the keywords in the instructions contain a large amount of information, which enables the keywords to represent the complete semantic information of the instructions in some cases. Therefore, highlighting some keywords in the classification vector can significantly improve the accuracy of multi-intention recognition [27,28]. We found that although the <cls> word vector of Ernie-Gram's output contains the semantic information of the instructions, the model assigns equal importance to each character or word in information extraction. This results in semantic information loss in information extraction. In addition, the dimension limitation of the classification vector also cause a loss of instruction semantic information. Starting from the goal of minimizing semantic information loss and improving the accuracy of model multi-intention recognition, a model of deep semantic information extraction using Ernie-Gram, BiGRU, and Attention is constructed based on the idea that CNN uses different filters for feature extraction. The results show that the proposed model can effectively extract the coarse and fine-grained semantic information, balance the contribution of each piece of coarse and fine-grained information, and significantly improve the accuracy of multi-intention recognition.

As shown in the details of architecture in Table 1, we first use the Ernie-Gram model to encode the whole sentence, and take out each vector obtained after encoding as the input of the BiGRU module. Secondly, the BiGRU module generates two outputs, which are the set of hidden states at each time step and the hidden state vector output at the last step. We splice the forward and backward final hidden state vector of the output of the last layer of BiGRU and extract them as inputs to the attention layer. Then, the attention layer calculates the attention vector by taking the input as the query vector and the hidden state of each time step output by BiGRU as the key vector and value vector. We fuse the calculated attention vector with the query vector through an FC to obtain the final attention vector. We splice the final attention vector with the <cls> classification vector output by the Ernie-Gram model to obtain the final feature vector containing sentence coarse and fine-grained information, and send the feature vector into the FC to realize the multi-intention classification of ATC instructions.

4. Experiments

4.1. Experimental Data

Chinese radiotelephony communication instructions have strict standards, which are composed of numbers, units, airline code, airline name, and other elements, together with the information carried by the instruction itself in a certain format, which constitutes the whole content of ATC instructions [29,30]. The ATC instructions studied in this paper are based on the standard radiotelephony communications for air traffic services (MH/T4014-2003) issued by the Civil Aviation Administration of China, covering the instructions at the flight stage, approach stage, take-off and landing stage, and other stages. Figure 7 shows the tree structure diagram of common ATC instructions and pilots' response instructions. Based on this tree structure diagram, we described the process of human intention inference and matching. From this figure, it can be seen that we infer ATC instruction intentions by integrating information from different keywords, which is the starting point of the multi-intention recognition model we constructed.

In the ATC instruction intention recognition dataset we use, the dataset labels include single intention and composite intention, with a total of 15 categories. There are eight types of single intentions in the datasets, namely, adjustment, transfer, positioning of aircraft, reminder, waiting, permission, reporting, and verification. There are seven types of composite intentions in the datasets, which include adjustment and positioning, adjustment and reporting, adjustment and reminder, adjustment and transfer, transfer and reminder, location and reminder, and permission and adjustment. Referring to the standard, we defined the intention of ATC instructions as follows. (1) Adjustment class: the adjustment class refers to when the instruction contains a requirement for the aircraft to adjust to or maintain a certain state, such as adjustment of aircraft height, speed, heading, joining the planned route, bias, adjustment of aircraft call sign, transponder code, and other information expressed in the ATC instructions. (2) Transfer class: the transfer class refers to the intention shown by an air traffic controller in an instruction to transfer the control of an aircraft currently under his or her jurisdiction to a controller in another seat or area. (3) Positioning of aircraft class: the instruction contains the intent of air traffic controllers to determine the position of aircraft through electronic devices, such as radar and radio apparatus. (4) Reminder class: the intention of an instruction that does not contain enforcement but does contain information that would avoid a flight collision. (5) Waiting class: the instruction contains a message requiring the aircraft to wait for further definite notice and proceed as planned until then. (6) Permission class: the instruction contains the intention to agree to a request made by the aircraft or to release permission information in advance of an inquiry. (7) Reporting class: the instruction contains the intention to require the pilot to report information relevant to the flight. (8) Verification class: the instruction contains the intention to ask for information in the content of the ATC instructions, or to confirm that the content of the inquiry from the pilot and other intention information contained in the verification class is invalid. Furthermore, the composite intention is defined as the combination of single instruction intentions. The single intentions in the composite intention are in no particular order. Instance analysis of the dataset is shown in Table 2.

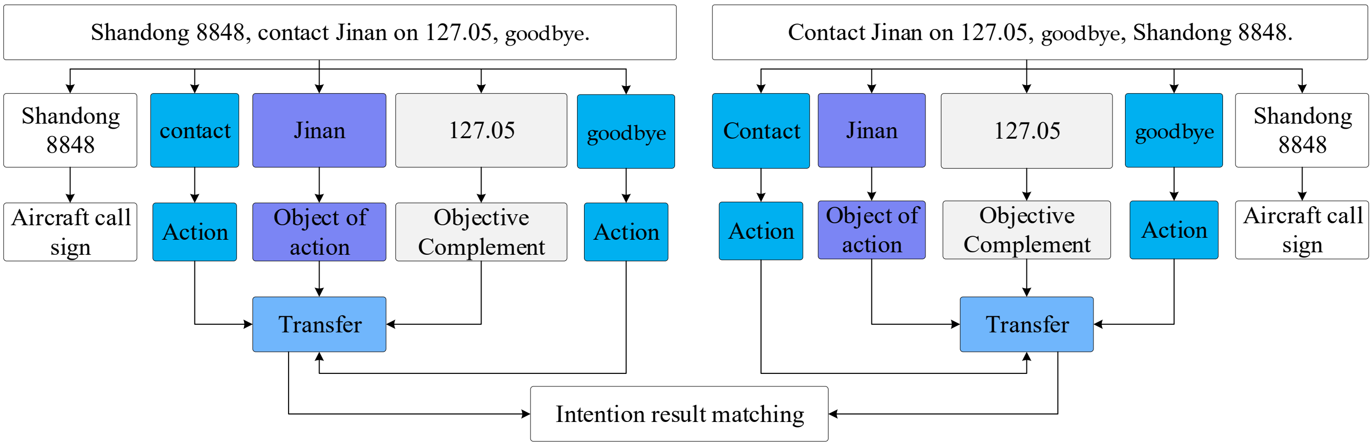


Figure 7. Instruction structure tree and intention recognition matching diagram.

Since there is no publicly available corpus of ATC instructions, we collected real speech data of radiotelephony communications. Based on the collected data, after our strict screening, we manually marked the data based on the above definition of intention. We obtained 9520 pieces of data that could be used for training. The statistical distribution of dataset labels is shown in Figure 8:

In Figure 8, the horizontal axis represents the intention category, and its specific meanings can be found in the legend; the vertical axis represents the quantity of categories.

Table 2. Instance analysis of the dataset.

Class of Intention	ATC Instructions	Instance Analysis
Adjustment	China Eastern 9916, veer right by five nautical miles; China Eastern 2453, climb to and maintain 5700; China Eastern 3401, reduce indicated airspeed to 255.	Aircraft transverse distance adjustment; aircraft altitude adjustment; aircraft speed adjustment.
Transfer	China Southern 0055, contact Zhengzhou on 130.0, good day.	Transfer of aircraft control.
Positioning of aircraft	Shunfeng 6954, Zhengzhou, radar contact.	The radar located the aircraft.
Reminder	China Eastern 2125, similar flight number Dongfang 2135, monitor closely.	Similar flight number conflict alert.
Adjustment and positioning	Lucky Air 9525, Zhengzhou, radar contact, climb to and maintain 7800.	The radar locates the aircraft and adjusts its altitude.
Waiting	Hainan 7188, maintain radio silence on this frequency, I'll call you later.	Aircraft waiting at communications channel.
Permission	Lucky Air 9970, push back and start up approved, taxiway alpha, runway 36R.	The controller granted the pilot's request for a slip-out.
Reporting	Okay Airways 6311, report at 25 nautical miles out, flying over Zhengzhou.	Fly over the Zhengzhou reporting point, report at 25 nautical miles from reporting point.
Adjustment and reporting	Yangtze River 5643, speed 260, report passing waypoint.	Aircraft speed adjustment and report past the reporting point.
Verification	China Southern 6960, affirm, adjust heading to 210.	Confirm course adjustment to 210.
Adjustment and reminder	Air China 1671, climb to 7200, there's a convergence ahead.	Aircraft altitude adjustment and conflict alert.
Adjustment and transfer	Shunfeng 5137, set QNH 5300, contact ground on 125.3.	Aircraft altimeter adjustment and control transfer.
Transfer and reminder	Hainan 3211, contact Shanghai Control on 120.8, and be aware of the passing time of the waypoint.	Control transfer and past report point time reminder.
Location and reminder	China Express 3211, radar contact, you can contact your own dispatch for communication.	Radar locates aircraft and reminds them to communicate with the flight despatcher first.
Permission and adjustment	Chongqing 1205, descend to 900, cleared for ILS approach, runway 03.	Adjust altitude and agree to use instrument approach procedure for aircraft approach.

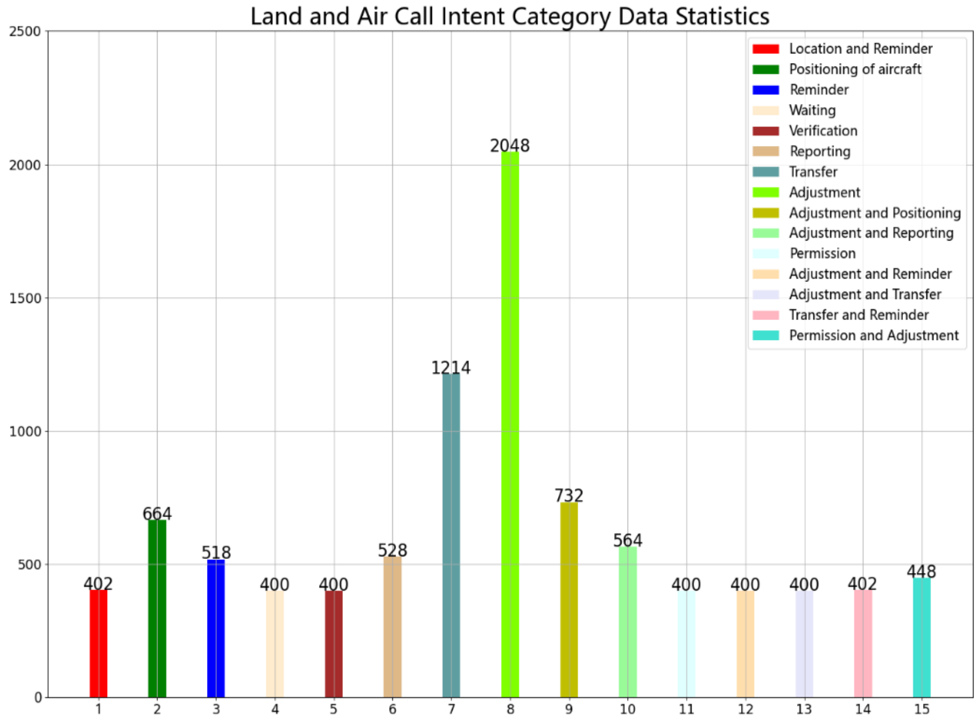


Figure 8. Label distribution of dataset.

4.2. Experimental Platform

Our experimental environment and configuration are as follows. Our experiment operating system is Windows10, CPU is an Intel(R) Xeon(R) E5-2680, and GPU is a NVIDIA RTX2080Ti 11 G; The Paddle framework was used to build the neural network model. The super parameter settings of the EBA model we constructed are shown in Table 3.

Table 3. EBA model hyperparameters table.

Hyperparameters	Number
Dropout	0.2
Max sequence length	50
Learning rate	2×10^{-5}
Batch size	32
Number of epochs	10

4.3. Ablation Experiment

We conducted ablation experiments to validate the effectiveness of each module in the constructed model and the feature fusion strategy adopted by the constructed model. The results are shown in Tables 4 and 5.

Table 4. Ablation experiment of EBA model.

Model	Dev_Acc	Recall	Precision	F1
EG	0.970	0.969	0.971	0.970
EGB	0.977	0.977	0.977	0.977
EBA	0.982	0.982	0.981	0.981

Table 5. Feature fusion strategy ablation experiment.

Feature Fusion Method	Dimensionality of the Fused Features	Dev_Acc	Recall	Precision	F1
Not fused (using only feature vectors from the base layer)	768	0.970	0.971	0.969	0.970
Not fused (using only feature vectors from the BiGRU Attention layer)	768	0.065	0.070	0.065	0.068
Dimensionality-changing feature fusion	768	0.058	0.062	0.058	0.060
	1536	0.077	0.080	0.077	0.079
Additive feature fusion	768	0.972	0.972	0.972	0.972
Concatenation feature fusion	1536	0.982	0.982	0.981	0.981

In Table 4, the abbreviation EG represents the Ernie-Gram model, EGB represents Ernie-Gram BiGRU, and EBA represents Ernie-Gram BiGRU Attention. As can be seen from this table, we can effectively improve the performance of the model by using the strategy of integrating the feature extraction of text information from different models.

From the table above, it can be seen that when we feed the feature vectors from the base layer output to the output layer, the model's accuracy can reach 0.97, whereas when we only use the feature vectors from the BiGRU Attention layer output, the model's accuracy is only 0.065. The reason for the poor performance of the latter is that the attention mechanism extracts local information for intent recognition, which cannot effectively reflect the intention of ATC instructions contained in the text. When we use feature fusion strategies, the model's accuracy and other performance indicators are further improved. For example, when using additive feature fusion strategy, the model's accuracy is improved to 0.972. When we also consider the feature dimensions in feature fusion, the model's performance is further improved. For instance, when using feature concatenation as the feature vector fusion method, the model's accuracy reaches 0.982. In conclusion, using an appropriate feature fusion method to combine the global information output by the base model with the local information output by Attention can significantly improve the performance of the model. For the multi-intention recognition of ATC instructions, using feature concatenation for feature fusion is a good method.

4.4. Experimental Analysis

We adopted the Ernie-Gram pre-trained language model as the bottom layer. In the training process, we did not update the pre-training model parameters, but only updated the model parameters of the BiGRU module, the Attention layer, and the FC. Through 10-epoch training, we obtained the experimental results shown in Table 6. The accuracy rate of each round of the model testing in the real process is drawn as a broken line graph in Figure 9. In terms if the evaluation criteria of the model, the precision rate, accuracy rate, F1 rate, and recall rate were adopted for comparative evaluation of the model [31]. To assess the model's performance improvement, we used a confusion matrix to visualize its effects. The formula for calculating indicators is as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$P = \frac{TP}{TP + FP} \tag{20}$$

$$F1 = \frac{2PR}{P + R} \tag{21}$$

Table 6. Comparison of the effect of different models on the ATC instruction intention recognition dataset.

Model	Dev_Acc	Recall	Precision	F1
BERT	0.957	0.956	0.957	0.956
BERT-Jieba tokenizer	0.911	0.912	0.911	0.911
BERT-LSTM	0.956	0.92	0.956	0.937
BERT-BiLSTM	0.951	0.93	0.951	0.94
BERT-TextCNN	0.96	0.96	0.97	0.96
BERT-GRU	0.968	0.967	0.968	0.967
BERT-BiGRU	0.971	0.969	0.971	0.97
BERT-TextCNN-BiGRU	0.975	0.974	0.975	0.974
EBA	0.982	0.982	0.981	0.981

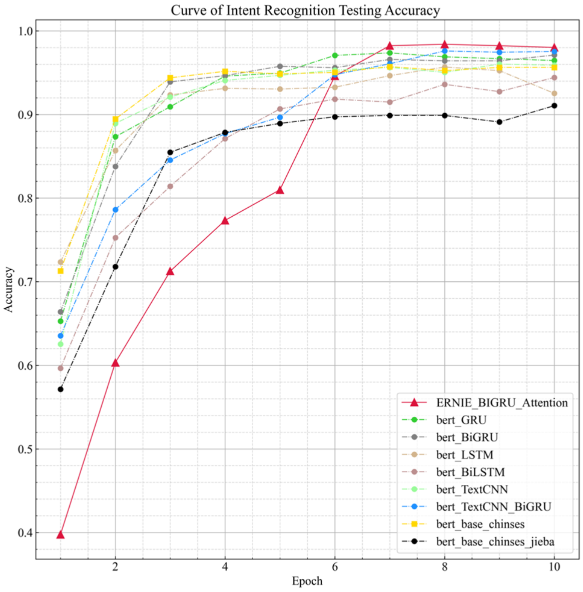


Figure 9. The accuracy curve of ATC instruction intention recognition on the test set.

P represents the precision rate, *Acc* represents the accuracy rate, *R* represents the recall rate, and *F*1 represents the harmonic average of precision and recall. *TP* represents the number of positive samples predicted by the model as a positive class, *TN* represents the number of negative samples predicted by the model as a negative class, *FP* represents the number of negative samples predicted by the model as a positive class, and *FN* represents the number of positive samples predicted by the model as a negative class.

From Figure 9, it can be seen that the EBA model has lower accuracy in the initial stage. However, with the increase in training rounds, the accuracy of the EBA model continues to improve in the fifth to sixth rounds, while other comparative models have already converged. Ultimately, the accuracy of the EBA model reached 0.982, showing excellent performance.

According to the experimental results in Table 6 and in combination with the results provided in Table 4, it can be observed that the accuracy of BERT is 0.957, while that of EG is as high as 0.97. Compared to BERT, BERT-GRU, BERT-LSTM, BERT-BilSTM, and BERT-TextCNN, EG exhibits better performance, indicating that selecting EG as the base layer for the model is a reasonable choice. Furthermore, based on the experimental results in Table 6, BiGRU is more effective in terms of feature extraction compared to the LSTM, BiLSTM, TextCNN, and GRU modules. It is worth noting that, combined with the results in Tables 4 and 6, the performance of the EGB model was improved from 0.977 to 0.982 by introducing the attention mechanism, while TextCNN only improved the performance of BERT-BiGRU from 0.971 to 0.975. Therefore, the attention mechanism is a more effective local feature extraction strategy in ATC instruction intention recognition. Finally, after improving the tokenizer based on the WordPiece strategy in BERT to the Jieba tokenizer based on the word strategy, the final model's performance decreased from 0.957 to 0.911, indicating that changing the encoding strategy may not be an effective way to improve model performance.

The attention visualization diagram of the Ernie-Gram module in the EBA model on the ATC instruction intention recognition test set is shown in Figure 10.

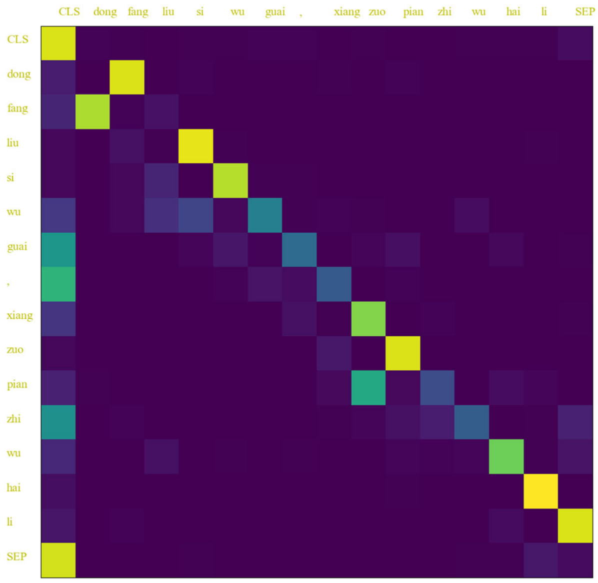


Figure 10. Heat map of attention results of Ernie-Gram module.

In Figure 10, the vertical axis represents the input ATC instruction text, and the horizontal axis represents the self-attention results, which indicate the degree of association between each word and other words in the context. The brighter the color, the deeper the association. The [CLS] vector is used to represent the semantic information of the entire input sequence. Based on the visualization results of Figure 10 and the experimental results, the high brightness of the [CLS] vector in the EG model indicates that the model pays more attention to the semantic information of the entire input sequence.

The visualization of attention results of the attention layer in the EBA model on the ATC instruction intention recognition test set is shown in Figure 11a,b, where Figure 11a is the visualization result for single intention test data and Figure 11b is the visualization result for multi-intention test data.

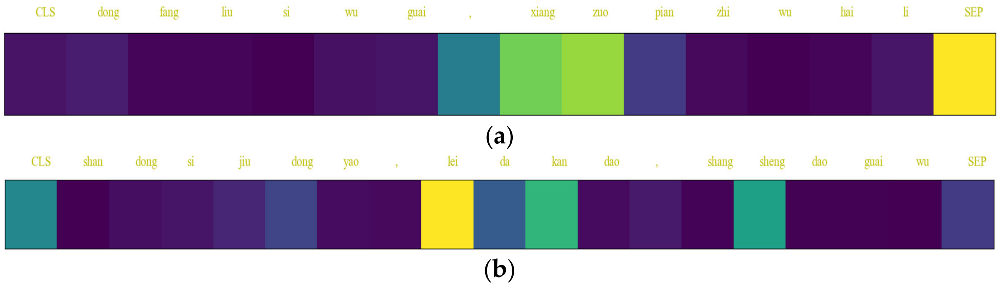
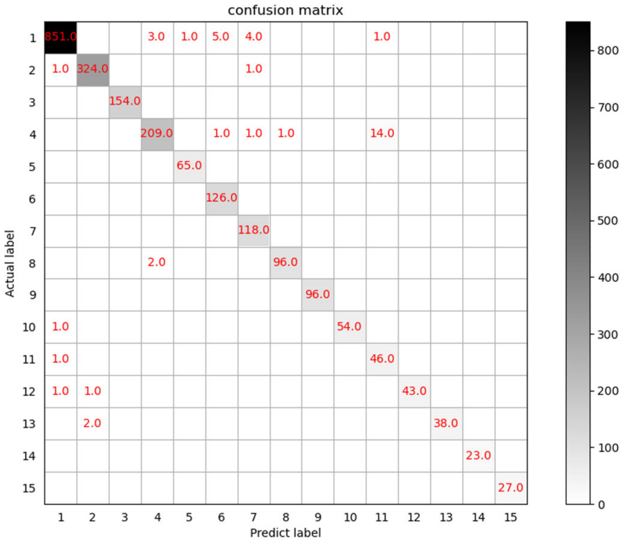


Figure 11. (**a**) Heat map of attention results of attention layer for single intention data. (**b**) Heat map of attention results of attention layer for multi-intention data.

In Figure 11, because the attention layer mainly focuses on local information in the input, we can see that in Figure 11a, the attention layer pays more attention to the words 'xiang', 'zuo', and 'sep', while in Figure 11b, the attention layer pays more attention to the words 'lei', 'da', 'kan', and 'sheng', all of which are critical keyword information for intention recognition, except for the character 'sep'. This indicates that the attention layer can effectively extract local keyword information in the ATC instruction intention recognition task.

The confusion matrix of test results of the EBA model on the ATC instruction intention recognition test set is shown in Figure 12.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	851.0			3.0	1.0	5.0	4.0				1.0				
2	1.0	324.0					1.0								
3			154.0												
4				209.0		1.0	1.0	1.0			14.0				
5					65.0										
6						126.0									
7							118.0								
8				2.0				96.0							
9									96.0						
10	1.0									54.0					
11	1.0										46.0				
12	1.0	1.0										43.0			
13		2.0											38.0		
14														23.0	
15															27.0

Figure 12. Test result confusion matrix of EBA model on the test set.

In this figure, both the horizontal and vertical coordinates are label categories, and the number at the position of the diagonal matrix represents the number correctly identified. As can be seen, the EBA model we constructed can achieve excellent results in the ATC instruction intention recognition dataset, especially for multi-intention data. The experimental results show that for multi-intention recognition, extracting semantic information from both global and local perspectives is an effective strategy to improve the accuracy of multi-intention recognition. In addition, the selection of a suitable feature fusion strategy and text encoding mode is also conducive to the improvement of model performance.

5. Conclusions

Based on the standard and the analysis of Chinese radiotelephony communication text data we collected, we extracted 15 common ATC instruction intentions, including 8 kinds of single intention and 7 kinds of composite intention. We compared and tested the effect of the Jieba tokenizer tool and the BERT tokenizer tool in the BERT model, and the result shows that the effect of the BERT tokenizer tool is superior to that of the Jieba tokenizer tool. We propose a multi-intention recognition model, EBA. The EBA model achieved 98.2% accuracy in the ATC intention identification dataset, which was 2.7% higher than that of the BERT benchmark model and 0.7% to 3% higher than that of other improved models based on BERT. In the future, due to the high safety requirements of ATC, our follow-up work will be devoted to the optimization of the model and the application of intention recognition results to the ATC human–computer dialogue system, so as to improve the accuracy of the machine to generate repeated response instructions according to the recognition intention.

Author Contributions: Conceptualization, W.P. and P.J.; methodology, W.P.; software, P.J.; validation, Z.W., W.P. and P.J.; formal analysis, Z.W.; investigation, Z.L.; resources, W.P.; writing—original draft preparation, W.P.; writing—review and editing, Z.W.; visualization, P.J.; supervision, Y.L.; project administration, W.P.; funding acquisition, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This study was co-supported by the National Key R&D Program of China (No. 2021YFF0603904), National Natural Science Foundation of China (U1733203), and Safety Capacity Building Project of Civil Aviation Administration of China (TM2019-16-1/3).

Data Availability Statement: Data are available on request due to restrictions, e.g., privacy or ethical concerns.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RNN	Recurrent neural networks
EBA	Ernie-Gram_BiGRU_Attention
BiGRU	Bidirectional gate recurrent unit
FC	Fully connected layer
ICAO	International Civil Aviation Organization
ATC	Air traffic control
ATS	Air traffic services
CAAC	Civil Aviation Administration of China
ATCO	Air traffic controller
SVM	Support vector machine
CNN	Convolutional neural networks
TextCNN	Convolutional neural networks for sentence classification
RCNN	Recurrent convolutional neural networks
LSTM	Long short-term memory
BiLSTM	Bidirectional long short-term memory
NLP	Natural language processing
NLU	Natural language understanding
MLM	Masked language model
EG	Ernie-Gram
EGB	Ernie-Gram bidirectional gate recurrent unit

References

1. Lin, Y. Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application. *Aerospace* **2021**, *8*, 65. [CrossRef]
2. Guo, S.; Wang, Q. Application of Knowledge Distillation Based on Transfer Learning of ERNIE Model in Intelligent Dialogue In-tention Recognition. *Sensors* **2022**, *22*, 1270. [CrossRef] [PubMed]
3. Liu, C.; Xu, X. AMFF: A new attention-based multi-feature fusion method for intention recognition. *Knowl. Based Syst.* **2021**, *233*, 107525. [CrossRef]
4. Dušek, O.; Jurčíček, F. A context-aware natural language generator for dialogue systems. *arXiv* **2016**, arXiv:1608.07076.
5. Haffner, P.; Tur, G.; Wright, J.H. Optimizing SVMs for complex call classification. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), Hong Kong, China, 6–10 April 2003.
6. Hakkani-Tur, D.; Tür, G.; Chotimongkol, A. Using semantic and syntactic graphs for call classification. In Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, Ann Arbor, MI, USA, 29 June 2005.
7. Kim, J.-K.; Tur, G.; Celikyilmaz, A.; Cao, B.; Wang, Y.-Y. Intent detection using semantically enriched word embeddings. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Juan, WA, USA, 12–13 December 2016.
8. Jeong, M.; Lee, G.G. Triangular-Chain Conditional Random Fields. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 1287–1302. [CrossRef]
9. Sutton, C.; McCallum, A. *An Introduction to Conditional Random Fields. Foundations and Trends® in Machine Learning*; IEEE: Picataway, NJ, USA, 2012; Volume 4, pp. 267–373.
10. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.
11. Wang, R.; Li, Z.; Cao, J.; Chen, T.; Wang, L. Convolutional Recurrent Neural Networks for Text Classification. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019.
12. Zhou, C.; Sun, C.; Liu, Z.; Lau, F. A C-LSTM Neural Network for Text Classification. *arXiv* **2015**, arXiv:1511.08630.
13. Liu, B.; Lane, I. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. *arXiv* **2016**, arXiv:1609.01454.
14. Lin, Z.; Feng, M.; Santos, C.N.D.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-Attentive Sentence Embedding. *arXiv* **2016**, arXiv:1703.03130.
15. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In Proceedings of the China National Conference on Chinese Com-putational Linguistics, Kunming, China, 18–20 October 2019.

16. Yepes, J.L.; Hwang, I.; Rotea, M. New Algorithms for Aircraft Intent Inference and Trajectory Prediction. *J. Guid. Control. Dyn.* **2007**, *30*, 370–382. [CrossRef]
17. Lin, Y.; Deng, L.; Chen, Z.; Wu, X.; Zhang, J.; Yang, B. A Real-Time ATC Safety Monitoring Framework Using a Deep Learning Approach. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, Auckland, New Zealand, 27–30 October 2019.
18. Zhang, S.; Kong, J.; Chen, C.; Li, Y.; Liang, H. Speech GAU: A Single Head Attention for Mandarin Speech Recognition for Air Traffic Control. *Aerospace* **2022**, *9*, 395. [CrossRef]
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Xiao, D.; Li, Y.-K.; Zhang, H.; Sun, Y.; Tian, H.; Wu, H.; Wang, H. ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding. *arXiv* **2020**, arXiv:2010.12148.
21. Lu, Q.; Zhu, Z.; Xu, F.; Zhang, D.; Wu, W.; Guo, Q. Bi-GRU Sentiment Classification for Chinese Based on Grammar Rules and BERT. *Int. J. Comput. Intell. Syst.* **2020**, *13*, 538–548. [CrossRef]
22. ArunKumar, K.; Kalaga, D.V.; Kumar, C.M.S.; Kawaji, M.; Brenza, T.M. Comparative analysis of Gated Recurrent Units (GRU), long Short-Term memory (LSTM) cells, autoregressive Integrated moving average (ARIMA), seasonal autoregressive Integrated moving average (SARIMA) for forecasting COVID-19 trends. *Alex. Eng. J.* **2022**, *61*, 7585–7603. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
24. Deng, J.; Cheng, L.; Wang, Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. *Comput. Speech Lang.* **2021**, *68*, 101182. [CrossRef]
25. Church, K.W.; Chen, Z.; Ma, Y. Emerging trends: A gentle introduction to fine-tuning. *Nat. Lang. Eng.* **2021**, *27*, 763–778. [CrossRef]
26. Cheng, X.; Zhang, C.; Li, Q. Improved Chinese Short Text Classification Method Based on ERNIE_BiGRU Model. In Proceedings of the 14th International Conference on Computer and Electrical Engineering (ICCEE), Beijing, China, 25–27 June 2021.
27. Zuluaga-Gomez, J.; Veselý, K.; Blatt, A.; Motlicek, P.; Klakow, D.; Tart, A.; Szöke, I.; Prasad, A.; Sarfjoo, S.; Kolčárek, P.; et al. Automatic Call Sign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications. *Multidiscip. Digit. Publ. Inst. Proc.* **2020**, *59*, 14.
28. Ding, X.; Mei, Y. Research on short text classification method based on semantic fusion and BiLSTM-CNN. In Proceedings of the 4th International Conference on Information Science, Electrical, and Automation Engineering (ISEAE 2022), Online, 25–27 May 2022.
29. Zhang, J.; Zhang, P.; Guo, D.; Zhou, Y.; Wu, Y.; Yang, B.; Lin, Y. Automatic repetition instruction generation for air traffic control training using multi-task learning with an improved copy network. *Knowl. Based Syst.* **2022**, *241*, 108232. [CrossRef]
30. Lin, Y.; Wu, Y.; Guo, D.; Zhang, P.; Yin, C.; Yang, B.; Zhang, J. A Deep Learning Framework of Autonomous Pilot Agent for Air Traffic Controller Training. *IEEE Trans. Hum. Mach. Syst.* **2021**, *51*, 442–450. [CrossRef]
31. Kici, D.; Malik, G.; Cevik, M.; Parikh, D.; Başar, A. A BERT-based transfer learning approach to text classification on software requirements specifications. In Proceedings of the Canadian Conference on AI, Vancouver, BC, Canada, 25–28 May 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.