Chinese Journal of Aeronautics, (2018), 31(12): 2280–2289

Chinese Society of Aeronautics and Astronautics
& Beihang University

Chinese Journal of Aeronautics

cja@buaa.edu.cn
www.sciencedirect.com

Intelligent checking model of Chinese radiotelephony read-backs in civil aviation air traffic control

Guimin JIA *, Fangyuan CHENG, Jinfeng YANG, Dan LI

Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China

Received 28 February 2018; revised 18 July 2018; accepted 12 September 2018
Available online 23 October 2018

KEYWORDS

Air traffic control;
Chinese radiotelephony read-backs;
LSTM;
Mean pooling;
MLP;
Semantic checking

Abstract Federal Aviation Administration (FAA) and NASA technical reports indicate that the misunderstanding in radiotelephony communications is a primary causal factor associated with operation errors, and a sizable proportion of operation errors lead to read-back errors. We introduce deep learning method to solve this problem and propose a new semantic checking model based on Long Short-Time Memory network (LSTM) for intelligent read-back error checking. A mean-pooling layer is added to the traditional LSTM, so as to utilize the information obtained by all the hidden activation vectors, and also to improve the robustness of the semantic vector extracted by LSTM. A MultiLayer Perceptron (MLP) layer, which can maintain the information of different regions in the concatenated vectors obtained by the mean-pooling layer, is applied instead of traditional similarity function in the new model to express the semantic similarity of the read-back pairs quantitatively. The K-Nearest Neighbor (KNN) classifier is used to verify whether the read-back pairs are consistent in semantics according to the output of MLP layer. Extensive experiments are conducted and the results show that the proposed model is more effective and more robust than the traditional checking model to verify the semantic consistency of read-backs automatically.

1. Introduction

Nowadays, radiotelephony communication still plays a very important role in civil aviation transportation which is the primary information transmitting mode for the air traffic controller and flight crew.[1] In order to make the aircraft operate safely and efficiently, the air traffic controller and flight crew must understand each other accurately. During the whole flight, the Air Traffic Controllers (ATCs) give different instructions to the fight crew by considering the flight phase, weather

* Corresponding author.
E-mail address: gmjia@cauc.edu.cn (G. JIA).

Peer review under responsibility of Editorial Committee of CJA.

Production and hosting by Elsevier

https://doi.org/10.1016/j.cja.2018.10.001
1000-9361

conditions, air traffic situations, etc. The flight crew also has to keep contact with different ATCs in different flight phases to confirm the received instructions, and report the flight status and any other necessary information related to flight. So exchanging the information correctly between the ATCs and the flight crew is a direct and vital factor for the air traffic safety.

Radiotelephony communication in civil aviation air traffic control is a kind of professional semi-artificial language,[2] which is built up based on natural languages. To improve the language norms and terminology standard in civil aviation communications, a series of diction and pronunciation standards or proposals have been formulated by the International Civil Aviation Organization (ICAO). The Civil Aviation Administration of China has also set up corresponding standards and requirements that ATCs and pilots have to obey. Although the civil aviation technologies have developed rapidly and communication equipment failures have decreased gradually during the past decades, unsafe aviation incidents related to radiotelephony communication errors, such as inaccurate read-backs, misunderstanding, non-standard diction, and incomplete content, still happen occasionally.[3,4] According to the investigation reports of NASA, more than fifty percent of the aviation incidents have relation to radiotelephony communication errors.[5] For example, Tenerife airport disaster in 1977, Urumchi air crash in 1993, and the crash of flight 965 in 1995 all had close relation to misunderstandings between ATCs and flight crew.

Many communication problems involve human errors of ATCs or pilots. In spite of the double check required in radiotelephony communication, these human errors are usually difficult to be aware of by ATCs and pilots themselves due to fatigue, tension, tremendous pressure, as well as fast speak speed and unchanged low intonation. With the rapid development of civil aviation transportation industry, controllers and pilots are usually working under enormous pressures. According to surveys, an ATC is usually in charge of more than eighty flights in one hour at the peak of traffic flow in the hub airports of China, which means that about ten flights need to be handled at the same time. So it is very likely to make mistakes in radiotelephony communication. Furthermore, pilots and ATCs do not always obey the radiotelephony communication standards strictly and completely in practice. Moreover, the studies of psycholinguistics have shown that in continuous speech perception the speakers will generate potential psychological expectation, which is helpful to enhance the listening comprehension but is also easy to cause ambiguity.

The data link applications in ATC system are developed to solve this problem partly by exchanging messages between pilots and air traffic controllers. At present, Pre-Departure Clearance (PDC) system and Data link-Automatic Terminal Information Service (D-ATIS), and Controller-Pilot Data Link (CPDL) are typical applications of data link in air traffic control and services. The main benefits of these applications are workload reduction for both pilots and air traffic controllers, as well as the increase in air traffic safety by reducing language and congestion-related communication errors. However, taking account of the construction of data link communication network in different routes and airports, the installation of specialized avionics and programs in aircraft, and the technology development of data link communication, the radiotelephony communication is still the primary mode in many countries, including China, and in most flight phases. In addition, many kinds of information have to be transmitted by air-ground data link system, such as position of aircraft, engine data, important flight information, and data exchange with airlines, etc. Moreover, the data link communication also grows rapidly and the channel capacity is approaching saturation in some hub airports. Data link communication and radiotelephony communication are usually designed to back up each other in practice. Therefore, detecting the radiotelephony communication errors automatically by computers instead of persons is essential and significant for civil aviation safety. This is the original motivation of our research.

To guarantee that the instructions are received accurately by pilots, read-backs are required according to the air traffic control standard, which means that the pilot reads the instructions back to the controller. It is one of the most important ways to ensure that the instructions are transmitted correctly. It is also crucial for the safety of civil aviation transportation. The common problems in civil aviation radiotelephony communications mentioned above usually result in read-back errors, and FAA technical reports indicate that miscommunications are historically identified as a primary causal factor associated with pilot operation errors. Although read-back errors do not cause pilots' operation errors directly, a sizable proportion of operation errors have relation to read-back errors.[6] At present, ATCs and flight crew usually make radiotelephony communication in Chinese for domestic flights in China, so the main purpose of this paper is to verify whether the Chinese read-backs are consistent in semantics with instructions automatically and intelligently. It has to be noted that, to integrate the proposed method into real ATC system, many scientific and technical problems regarding radiotelephony communications have to be studied furtherly, including continuous speech recognition of radiotelephony communications in real work environment, robust semantic expression, semantic analysis, and situation awareness of their conversations, etc. It involves many fundamental problems in speech recognition and natural language processing. Our research in this paper provides a novel method of semantic matching and verification of the instructions and read-backs, and it is also useful to address other natural language processing problems mentioned above.

Our work is mainly related to the sentence pair modeling based on deep learning. It is an important task to appropriately model the relation between two sentences or texts in many fields of information search and natural language understanding, such as paraphrase identification, query suggestion, machine translation, image retrieval, automatic summarization and semantic textual similarity.[7] Many algorithms have been proposed for semantic matching which is mainly based on lexical matching,[8,9] linguistic analysis,[10] and neural networks.[11–16] In recent years, some algorithms based on deep neural networks have achieved encouraging results in semantic analysis and matching by using word embedding methods to express the words meaning as vectors in semantic space,[17–23] which requires no prior knowledge of natural language and no external resource of structured semantic information.

The Recurrent Neural Network (RNN) is a special case of the recursive network in which the structure is simple linear chain. Not only the RNN is primarily used as a language model, but also it has been proved efficient in terms of constructing sentence representations with a linear structure.[17] RNN applies a hidden state to represent sentences, and repeatedly feeds the hidden state and word embeddings to the network to update the sentence representations.[24] RNN may suffer from gradient vanishing and exploding problems which limit the length of reachable context. RNN with Long Short-Time Memory network (LSTM)[25,26] unit solves such problems by introducing a memory cell and gates into the network. The Convolutional Neural Network (CNN) has also been modified to model the sentence meaning.[27,28] These developments provide extremely valuable references for us to study the semantic checking problem of Chinese read-backs in civil aviation radiotelephony communication.

In this paper, we propose a novel scheme based on LSTM to address the problem of Chinese read-back verification. A two-channel structure is built up to process the ATC's instructions and the pilot's read-backs, respectively. The word embedding method of one-hot vector is used to represent the segmented words of these sentence pairs for the length is usually much shorter and the relation of words is weaker than many other sentence modeling tasks. The traditional LSTM is also improved by adding a mean-pooling layer for the reason that all the words in the read-backs usually have important information related to the flight in practice. All the hidden activation vectors are combined in concatenation by the mean-pooling layer to fully utilize the semantics underlying different sources and also to balance the semantic representation of the former words and the latter words in a sentence. So the semantics extracted by the new model is more robust than traditional LSTM. A MultiLayer Perceptron (MLP) layer is used instead of traditional cosine similarity function to represent the relatedness of the sentence vectors of read-back pairs output by the pooling layer. Then the quantitative similarities of read-back pairs are put into the K-Nearest Neighbor (KNN) classifier to check whether the sentence pair is consistent in semantics. Extensive experiments are conducted based on homemade corpus of Chinese radiotelephony communications in civil aviation. The experimental results show that our new model proposed in this paper is more effective and more robust than the traditional checking model described in our previous work.

2. Semantic representation models

2.1. RNN model

RNN is a natural feed-forward neural network. The structure of RNN is similar to traditional neural network, which is composed of input layer, hidden layer and output layer. The difference among them is that RNN's hidden units are connected so that it can retain the previous information. The hidden unit of RNN is affected by both the current input sequence and the previous hidden units. RNN's propagation is related to time-step.[29,30] The unfolding structure of RNN with time is shown in Fig. 1.

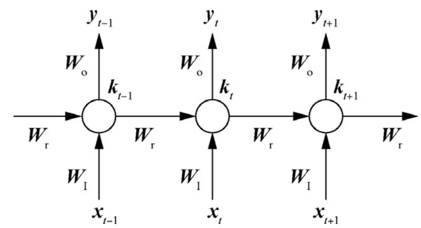


Fig. 1 Unfolding structure of RNN with time.

In Fig. 1, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T]$ is the input sequence, and $\boldsymbol{x}_t$ is the word vector of the t-th word. $\boldsymbol{K} = [\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_t, \ldots, \boldsymbol{k}_T]$ is the hidden sequence of the semantic model, and $\boldsymbol{k}_t$ is the hidden activation vector in time-step t. $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_T]$ is the output sequence. $\boldsymbol{W}_{\mathrm{I}}$, $\boldsymbol{W}_{\mathrm{o}}$ and $\boldsymbol{W}_{\mathrm{r}}$ are the input weight matrix, output weight matrix and recurrent weight matrix, respectively. RNN computes the hidden activation vector and output vector at time-step t by the following mathematical formulation:

$$\boldsymbol{k}_t = \tanh(\boldsymbol{W}_{\mathrm{I}}\boldsymbol{x}_t + \boldsymbol{W}_{\mathrm{r}}\boldsymbol{k}_{t-1} + \boldsymbol{b}) \tag{1}$$

$$\boldsymbol{y}_t = \sigma(\boldsymbol{W}_{\mathrm{o}}\boldsymbol{k}_t) \tag{2}$$

where $\boldsymbol{b}$ is the bias and we set bias as zero in our work; $\sigma(\cdot)$ is the sigmoid function. We use $\tanh(\cdot)$ and $\sigma(\cdot)$ as the activation function of hidden layer and output layer, respectively. Though it is theoretical that RNN is able to process arbitrary length sequences, it tends to suffer from the gradient vanishing and exploding problems which limit the length of reachable context.

2.2. LSTM model

In order to learn long-term memory of the input sentences and to address such problems of RNN, the LSTM unit is incorporated into the RNN model.[31] As shown in Fig. 2, the standard LSTM unit contains output gate $\boldsymbol{o}_t$, forget gate $\boldsymbol{f}_t$, input gate $\boldsymbol{i}_t$, memory cell $\boldsymbol{c}_t$ and hidden state $\boldsymbol{h}_t$.[32,33] The role of the gates is to control the flow of information, that is, which information should be input, forgotten and output, respectively. The output gate is used to shield error by taking itself offline, the forget gate is used to empty the memory contents, and the input gate is used to ignore incoming activations. The memory cell is the core of LSTM unit to store information over long-time durations. $\boldsymbol{h}_t$ is regarded as a representation of the current time step with contexts. In order to generate $\boldsymbol{h}_t$, a temporary result $\boldsymbol{l}_t$ is first computed by tanh over the ensemble of input $\boldsymbol{x}_t$ and the preceding hidden state $\boldsymbol{h}_{t-1}$.

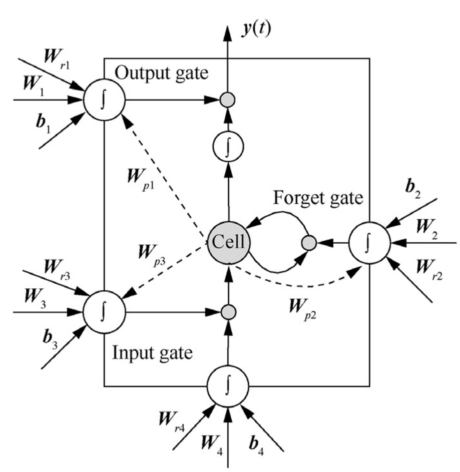


Fig. 2 Architecture of LSTM.

In this paper, the LSTM unit contains peephole connections to fully utilize the presentation of memory cell. The transition functions of LSTM are shown as Eqs. (3)–(8).

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_3\boldsymbol{x}_t + \boldsymbol{W}_{r3}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{p3}\boldsymbol{c}_{t-1}) \tag{3}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_2\boldsymbol{x}_t + \boldsymbol{W}_{r2}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{p2}\boldsymbol{c}_{t-1}) \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_1\boldsymbol{x}_t + \boldsymbol{W}_{r1}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{p1}\boldsymbol{c}_{t-1}) \tag{5}$$

$$\boldsymbol{l}_t = \tanh(\boldsymbol{W}_4\boldsymbol{x}_t + \boldsymbol{W}_{r4}\boldsymbol{h}_{t-1}) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t\boldsymbol{l}_t \tag{7}$$

$$\boldsymbol{h}_t = \tanh(\boldsymbol{c}_t)\boldsymbol{o}_t \tag{8}$$

where $\boldsymbol{W}_i$ and $\boldsymbol{W}_{ri}$ (r, i = 1, 2, 3, 4) are input connections and recurrent connections of output gate, forget gate, input gate and cell, respectively; $\boldsymbol{W}_{pi}$ (p, i = 1, 2, 3) are peephole connections; $\tanh(\cdot)$ and $\sigma(\cdot)$ are activation functions.

3. Proposed checking model

A new model is proposed to solve the checking problems of Chinese radiotelephony read-backs, as illustrated in Fig. 3. The ATC's instructions and the pilot's read-backs have been transformed into text format, and the semantics of the sentences are extracted and verified by this scheme which has two channels to process the sentences and to learn the semantics of instructions and read-backs, respectively. The procedure of the new model is elaborated as follows:

Step 1. The instructions ($S_{\rm ATC}$) and the read-backs ($S_{\rm P}$), which have been converted into text format, are segmented into single words by the Chinese word segmentation method.

Step 2. The words of $S_{\rm ATC}$ and $S_{\rm P}$ are mapped into vector space by word embedding methods. According to our previous study,[34] the sentence length of read-backs is usually shorter and the relation among the words is weaker than common sentence modeling tasks, so each segmented word is coded with one-hot vector in this paper.

Step 3. The semantic model of $S_{\rm ATC}$ ($M_{\rm ATC}$) and the semantic model of $S_{\rm P}$ ($M_{\rm P}$) are established based on LSTM, respectively. The semantic representation models $M_{\rm ATC}$ and $M_{\rm P}$ have the same structure. They both have two layers which are input layer and hidden layer. The input layer is used to input segmented word vectors of $S_{\rm ATC}$ or $S_{\rm P}$. The hidden layer is applied to generate the semantic vector of the whole sentence. The model of $M_{\rm ATC}$ or $M_{\rm P}$ activates one input word vector for each time-step in sequence, which may obtain the contextual information of sentence.[35]

Supposing that $\boldsymbol{h}_t$ is the hidden activation vector in time-step t, it represents the semantic information entered at time-step t and before. When the last word vector $\boldsymbol{x}_T$ of the sentence is input into the semantic model, the corresponding hidden activation vector $\boldsymbol{h}_T$ is supposed to be the semantic representation vector of the whole sentence,[29] and it puts more emphasis on the latter words.

Step 4. To represent the whole sentence more properly and express each word more equally, at the last time-step T, the output of LSTM model is put into a mean-pooling layer, whose structure is illustrated in Fig. 4.

The pooling layer can help a model retain the most prominent and prevalent features, which is helpful for robustness across examples. In this layer, all the outputs of hidden layer of LSTM are combined in a concatenation as the input of the pooling layer. The widely adopted mean pooling is applied and the output of mean pooling is the average vector of all outputs of hidden layer of LSTM. For the hidden activation vector in time-step t is the semantics of the t-th word, the mean-pooling layer can fully utilize the semantics of all words in $S_{\rm ATC}$ or $S_{\rm P}$. The semantics of the former words are also balanced against the semantics of the latter ones. So the output of this layer is the feature representation of sentence $S_{\rm ATC}$ ($\boldsymbol{v}_{\rm ATC}$), or the feature representation of sentence $S_{\rm P}$ ($\boldsymbol{v}_{\rm P}$).

Step 5. The semantic difference between $S_{\rm ATC}$ and $S_{\rm P}$ is transformed into the distance between $\boldsymbol{v}_{\rm ATC}$ and $\boldsymbol{v}_{\rm P}$ in the vector space. A MLP layer is adopted to learn and extract the semantic similarity quantitatively. In this part, MLP is constituted by input layer, hidden layer and output layer. $\boldsymbol{v}_{\rm ATC}$ and $\boldsymbol{v}_{\rm P}$ are concatenated into a vector, which is the input of the MLP layer. The output of MLP is used to indicate the semantic relatedness of $S_{\rm ATC}$ and $S_{\rm P}$.

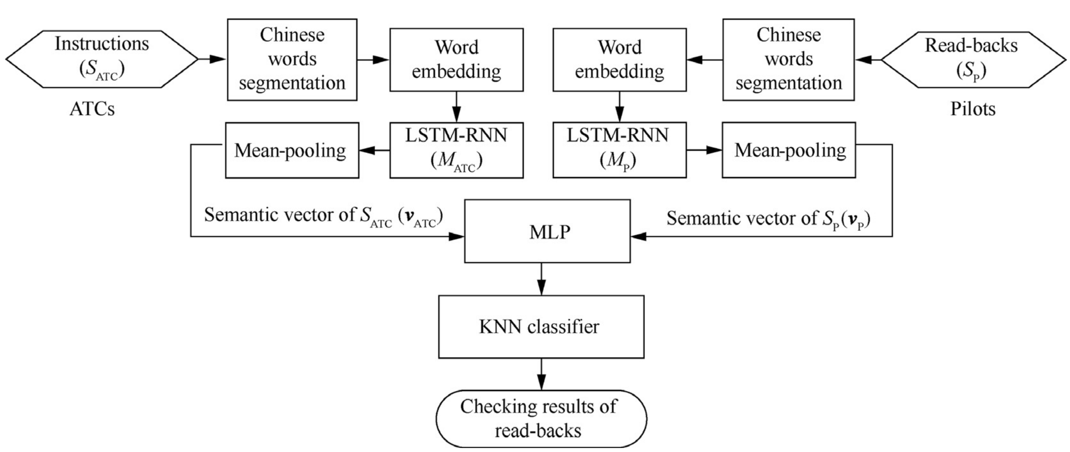


Fig. 3 Flowchart of proposed checking scheme.

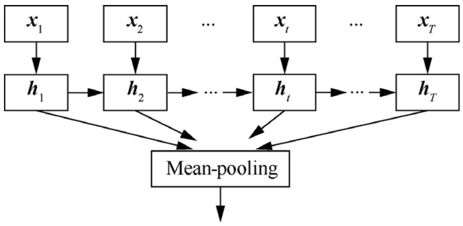


Fig. 4 LSTM model with mean-pooling layer.

Step 6. A classifier layer is applied to classify the output of MLP layer into two classes so as to verify whether the semantics of S_{ATC} and S_{P} is consistent or not. According to our previous study,[34] KNN is used in our work.

The two models of M_{ATC} and M_{P} have to be trained together according to the value and the variance of similarity of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$ in the training process. The corpus used in our work is labelled, and then supervised learning method is utilized to train the model. The training processes of M_{ATC} and M_{P} are the same. We train the models by minimizing the cross-entropy error. The error is defined as

$$J = -\sum_{n=1}^{N}[L\lg S_{\mathrm{r}} + (1-L)\lg(1-S_{\mathrm{r}})] \tag{9}$$

where S_{r} is the semantic relatedness which is the output of MLP layer in our checking model. L is the labelled number, and if the sentence pair is consistent in semantics, L is defined as one, otherwise L is defined as zero. We use Back Propagation Though Time (BPTT) to learn the model parameters and the momentum method is used to accelerate the convergence.[36]

4. Experiments and analysis

4.1. Corpus of Chinese ATC radiotelephony read-backs

In order to analyze the performance of the new checking model for the consistency verification of ATC sentence pairs, we built an experimental civil aviation radiotelephony communication corpus (ATC Corpus) in Chinese. The corpus is built up according to the original recordings of radiotelephony communication between pilots and air traffic controllers, and also the professional books used for civil aviation radiotelephony training. Since there are lots of proper nouns in the recordings, the pronunciation of numbers and letters is different to common Chinese language, and the actual speed of ATC-pilot communication is much faster than common conversation. We invited two professional air traffic controllers to listen to these recordings and translated them to text format separately so as to ensure the correctness of the corpus.

In this paper, we chose sentence pairs of read-backs ($S_{\mathrm{ATC\text{-}P}}$) from the corpus to conduct these experiments. 800 $S_{\mathrm{ATC\text{-}P}}$ pairs that are annotated with consistency (the positive samples) were selected from the corpus. For improper read-backs do not occur regularly, we chose and designed different types of improper read-backs according to the investigation of communication problems in civil aviation. 500 $S_{\mathrm{ATC\text{-}P}}$ pairs that are inconsistent in semantics formed the negative samples, so the total samples are 1300 $S_{\mathrm{ATC\text{-}P}}$ pairs that have covered the common ATC call codes and the conversations during the whole flight phase. The train set consists of 500 positive samples and 300 negative samples, which are picked up randomly from the total samples. The test set (Total) consists of the other rest samples, in which 300 samples are positive (Positive) and 200 samples are negative (Negative).

4.2. Experimental results based on traditional checking model

According to the traditional RNN and LSTM, a traditional checking model of Chinese read-backs was proposed in Ref. [35]. To compare the proposed checking model, the experiments based on traditional model were conducted. The procedure of traditional checking model is shown in Fig. 5.

The preprocessing of the $S_{\mathrm{ATC\text{-}P}}$ pairs is the same, while the outputs of RNN/LSTM models are directly put into the cosine similarity layer to compute the relatedness of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$. The KNN classifier is applied to verify the semantic consistency of sentence pair. The one-hot vector was used to represent segmented words of $S_{\mathrm{ATC\text{-}P}}$ pairs. The two channels are trained together according to the variance of cosine similarities of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$.

Similarly, test accuracy and Mean Squared Error (MSE) are used to evaluate the performance of the checking model. For the checking of read-backs is of crucial importance to the safety of aviation transportation, the checking model has to be more sensitive to improper read-backs. So in the experiments, the test accuracy of Total (P_{t}), the test accuracy of Positive (P_{p}), and the test accuracy of Negative (P_{n}) are computed and compared. The experimental results based on the data set described in Section 4.1 are shown in Fig. 6 and Table 1.

In the experiments, the model was trained and tested thirty times randomly. Fig. 6 illustrates the test accuracies based on RNN and LSTM. From Fig. 6 and Table 1, we can figure out that the test accuracy based on LSTM model is more stable than that based on RNN model, and the traditional checking model based on LSTM achieves the highest average accuracy in Positive, Negative, and Total. The average test accuracy of Positive is higher than that of Negative in both models of

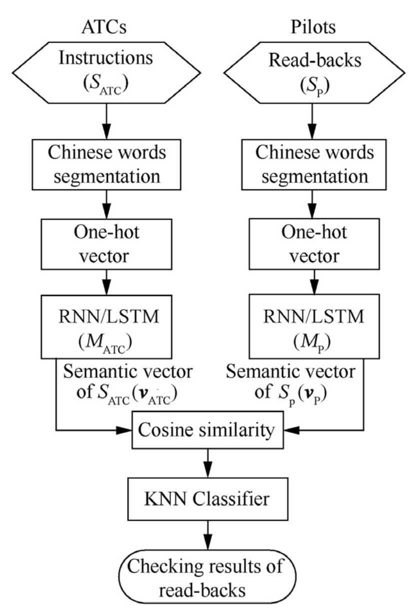


Fig. 5 Traditional checking model of Chinese read-backs.

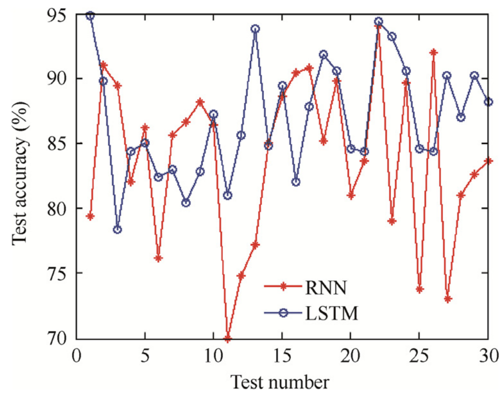


Fig. 6 Test accuracy of traditional checking model.

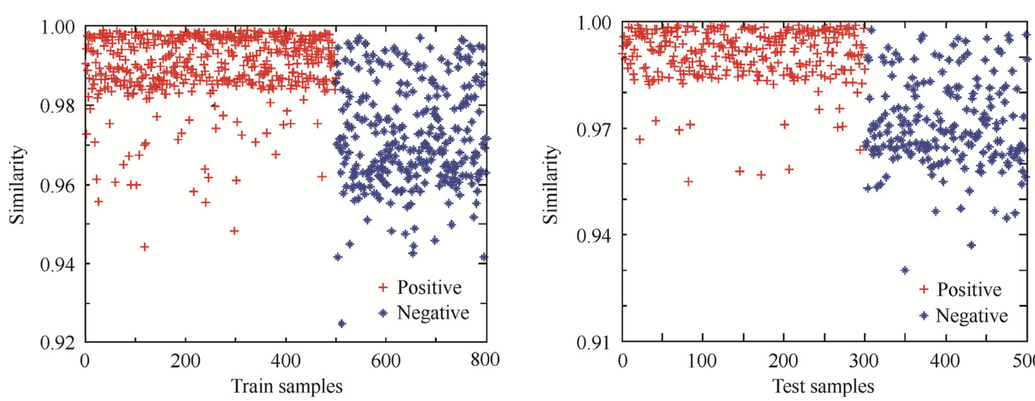
Table 1 Average test accuracy and MSE of traditional checking models.

Model	Test accuracy (%)			MSE
	P_p	P_n	P_t	
RNN	88.42	76.97	83.84	0.062
LSTM	90.02	82.20	86.89	0.037

RNN and LSTM. Besides, Figs. 7 and 8 show the distribution of similarities in one random test of traditional checking model based on RNN and LSTM, respectively. It can be seen that all the similarities acquired by the traditional model based on RNN are very high. But in fact the negative sample's similarity should be lower than the positive sample's similarity. There should be obvious distinctions between the positive samples and the negative samples if their semantics is expressed properly enough by the model. By contrast, the similarities acquired by the traditional model based on LSTM have more obvious distinctions. It is indicated that LSTM has more advantages in semantic extraction. So the experimental results give a baseline to be compared with the proposed checking model of read-backs in this paper.

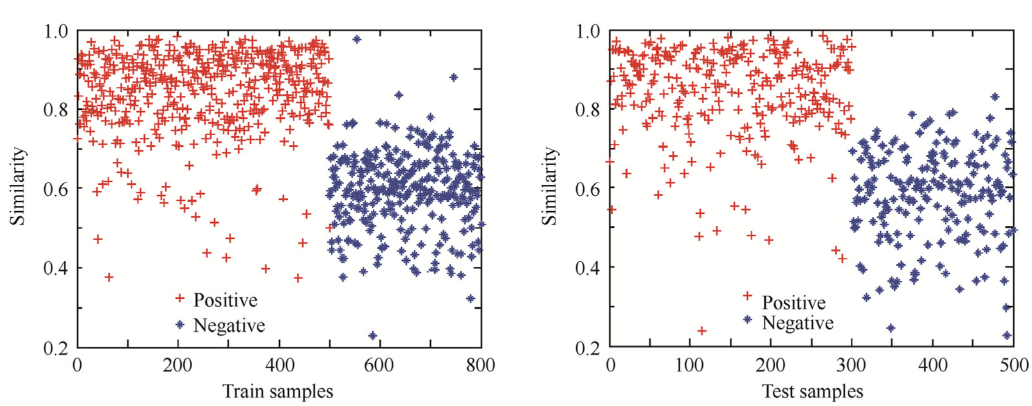
4.3. Experimental results of proposed checking model

To analyze the performance of the new checking model with added mean-pooling layer and MLP layer, the experiments are conducted as follows: (A) only the mean-pooling layer is added to the checking model based on LSTM, and the similarity of $\boldsymbol{v}_{ATC}$ and $\boldsymbol{v}_{P}$ is also computed by cosine similarity; (B) only the MLP layer is added to the checking model based on LSTM, and the outputs of hidden layer at the last time-step T are concatenated into a vector, as shown in Fig. 9, where



(a) Train samples (b) Test samples

Fig. 7 Distribution of similarities of traditional checking model based on RNN.



(a) Train samples (b) Test samples

Fig. 8 Distribution of similarities of traditional checking model based on LSTM.

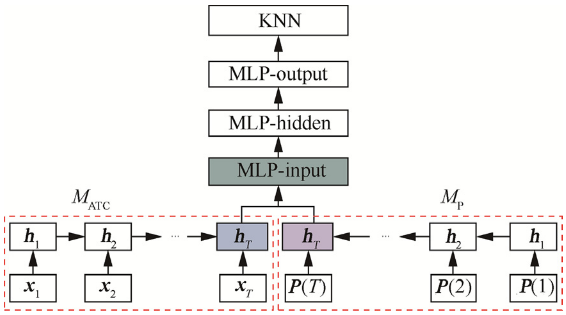


Fig. 9 LSTM model with MLP.

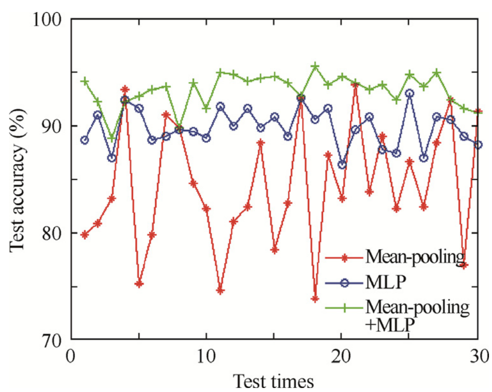


Fig. 10 Test accuracy of checking model with different layers based on LSTM.

[$\boldsymbol{P}(1)$, $\boldsymbol{P}(2)$, . . ., $\boldsymbol{P}(T)$] is the input sequence of pilot's read-back; (C) the proposed model with both mean-pooling layer and MLP layer is tested.

The experimental procedure is similar to the traditional checking model. The experiments with different layers are conducted separately, and the data set described in Section 4.1 is used during the experiments. Thirty independent tests are arranged for each experiment, and each test is conducted with randomly chosen train set and test set from the total samples.

The test accuracies of the checking model with different layers based on LSTM are illustrated in Fig. 10. The average test accuracy and MSE are shown in Table 2. As indicated in Fig. 10, the test accuracy of the proposed model (Mean-pooling + MLP) is obviously higher than that of the other two models with different layers, and the model with MLP layer achieves the secondary test accuracy. The fluctuation of the proposed model is the smallest in the experiments.

From Table 2, we can see that the average test accuracy of Negative is obviously improved in the three experiments compared to the traditional checking model by introducing different layers to the LSTM model. Also, it can be seen that there is a better balance among the average test accuracy of Positive and Negative by adding the pooling layer or/and the MLP layer. That is, the added pooling layer and MLP layer are helpful to improve the performance of the checking model.

The checking model based on LSTM with MLP layer has lower MSE and higher average accuracy than the traditional checking model based on LSTM. It indicates that the relatedness of the semantic vectors ($\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$) is expressed more

Table 2 Average test accuracy and MSE based on proposed model.

Model with different layers	Test accuracy (%)			MSE
	P_{p}	P_{n}	P_{t}	
Mean-pooling	84.94	83.52	84.37	0.056
MLP	89.94	89.62	89.81	0.017
Proposed model (Mean-pooling + MLP)	94.10	92.74	93.29	0.015

accurately by MLP instead of similarity function, and more interaction of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$ are taken into account by this layer.

However, the average test accuracy of the checking model with mean-pooling layer is lower than that of the traditional checking model. The primary reason is that the cosine similarity after the mean-pooling layer discards some important information because the particular regions of concatenated vectors, which are obtained by the mean-pooling layer, come from different underlying sources. Moreover, by comparing the experimental results of the proposed model and the model with MLP layer, it can be indicated that by concatenating all the hidden activation vectors the mean-pooling layer introduces more information. The features extracted after mean-pooling layer are more prevalent. So it is easier to learn the relatedness of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$ for the latter MLP layer.

It can be seen that all the average accuracies of the Total, Positive and Negative have been increased in the experiment, and at the same time the MSE has been reduced, which means that the proposed model is more robust and more superior for the checking task of Chinese read-backs. Figs. 11–13 show the distribution of similarities of total samples in one random test based on LSTM with different layers, respectively. It can be apparently seen that the similarities have better discrimination than that obtained by traditional model, and the new model has the best distinction in the similarity of $\boldsymbol{v}_{\mathrm{ATC}}$ and $\boldsymbol{v}_{\mathrm{P}}$. From the analysis above, we can figure out that by adding a mean-pooling layer and a MLP layer to the LSTM the new model is more suitable to extract the semantics of $\mathrm{S}_{\mathrm{ATC\text{-}P}}$ pairs and to represent the similarity relatedness of the $S_{\mathrm{ATC\text{-}P}}$ pairs.

4.4. Comparison results in extended ATC Corpus

In order to make further test of the proposed model, we extended the ATC Corpus according to real recordings and professional training books of radiotelephony communications. There are 2442 $S_{\mathrm{ATC\text{-}P}}$ pairs in the extended ATC Corpus, including 1326 positive samples (consistent in semantics) and 1116 negative samples (inconsistent in semantics). The traditional checking models based on RNN and LSTM are compared to the proposed model in this section. All the experimental results shown in this section are conducted on the extended ATC Corpus. For each test, according to the experimental procedure described in Section 4.1, the train set consists of 1952 randomly selected samples, in which there are 1060 positive samples and 892 negative samples. The other 490 samples comprise the test set, in which 266 samples are positive, and 224 samples are negative. Each model was trained and tested 30 times randomly. The test accuracy and MSE are calculated and compared, as shown in Fig. 14 and Table 3.

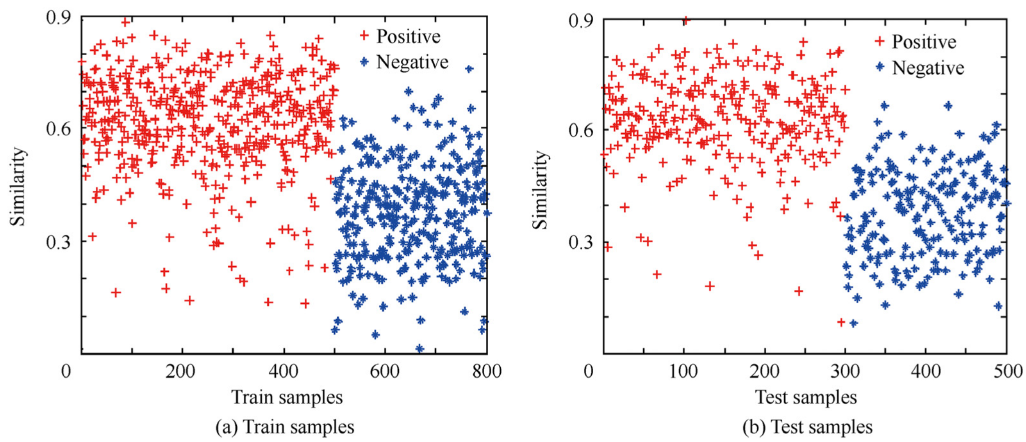


Fig. 11 Distribution of similarities of total samples based on LSTM with mean-pooling layer.

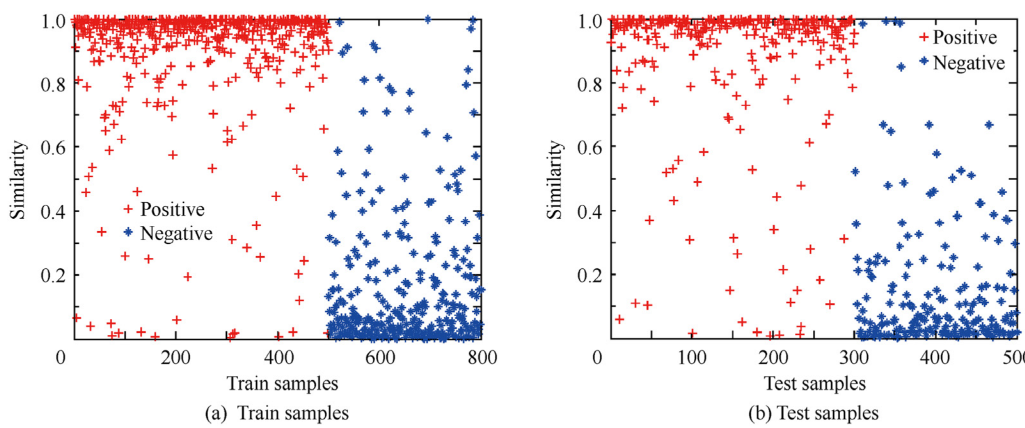


Fig. 12 Distribution of similarities of total samples based on LSTM with MLP layer.

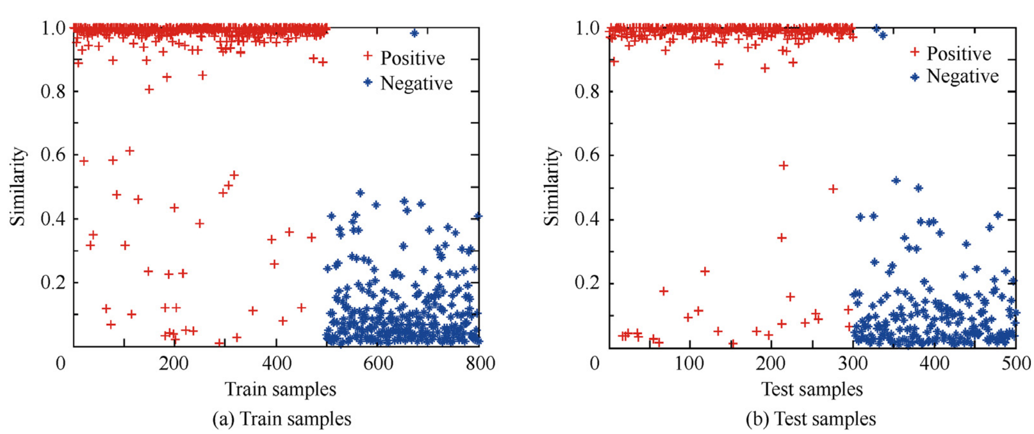


Fig. 13 Distribution of similarities of total samples based on proposed model.

It can be seen that the proposed model in this paper has greatly improved the performance in read-backs checking of radiotelephony communications compared to traditional checking model. The average test accuracies of Total, Positive, and Negative are also improved obviously, and the average test accuracy of Total test set is raised up to 92.47%. The average test accuracy of Negative has been increased to 90.82% which means that the new model makes a better balance

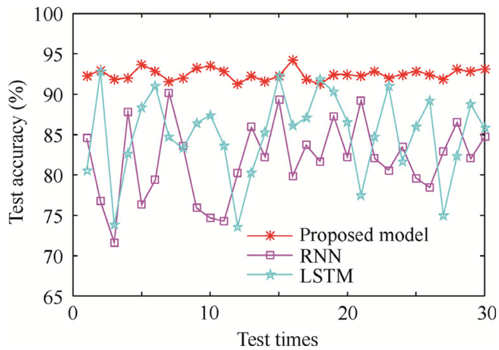


Fig. 14 Test accuracy of different models in extended ATC Corpus.

Table 3 Comparison results of average test accuracy and MSE in extended ATC Corpus.

Model	Test accuracy (%)			MSE
	P_p	P_n	P_t	
RNN	85.56	78.76	81.90	0.047
LSTM	87.18	82.70	85.00	0.053
Proposed model	93.77	90.82	92.47	0.007

among the average test accuracies of Positive and Negative samples. Therefore, the new model is more sensitive to the read-back errors. As shown in Fig. 14 and Table 3, the performance of the new model is much more stable than that of traditional checking models based on RNN and LSTM, and the MSE has been reduced obviously. So, as analyzed in Section 4.3, the comparison results in the extended ATC Corpus have further proved the outstanding performance of the proposed model for Chinese read-back checking.

5. Conclusions

To check the radiotelephony read-backs in civil aviation ATC intelligently and automatically, a proposed model is proposed based on LSTM. According to the experimental results and analysis described above, we can make conclusions as follows:

(1) The deep learning method is introduced to model the semantics of the radiotelephony communications in ATC, and a new semantic checking model of read-back errors is proposed in this paper to verify whether the meanings of the $S_{\text{ATC-P}}$ pairs are consistent.

(2) The traditional LSTM is improved by adding a mean-pooling layer to concatenate all the hidden activation vectors. It has advantages in fully utilizing the semantic information underlying different sources. In addition, the mean-pooling layer is helpful to achieve a better balance among the semantic representation of the former words and the latter words. So the semantic feature of the whole sentence is more prevalent than that extracted by the traditional LSTM.

(3) The MLP layer supplies more interaction of the semantic vectors of $S_{\text{ATC-P}}$ pairs compared to traditional cosine similarity function. It can express the deep relatedness of the $S_{\text{ATC-P}}$ pairs more accurately, and as a result the differences of the quantitative similarities acquired by the MLP layer are more distinctive between the negative samples and the positive samples.

(4) The average test accuracy of Negative samples is increased significantly, and the performance of the proposed model is highly improved in terms of test accuracy and MSE. So, by adding a mean-pooling layer and a MLP layer, the proposed checking model is more robust to check the read-back errors intelligently and automatically. The proposed model described in this paper also gives an encouraging way to model the semantics of Chinese radiotelephony communications in civil aviation air traffic control.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Nos. 61502498, U1433120 and 61806208) and the Fundamental Research Funds for the Central Universities, China (No. 3122017001).

References

1. Veronika PO, Morrow DG. Improving pilot/air traffic control voice communication in general aviation. *Int J Aviat Psychol* 1998;**12**(4):341–57.
2. National Transportation Safety Board. Review of U.S. civil aviation accident, calendar year 2010. Washington, D.C.: National Transportation Safety Board. Report No.: Annual Review NTSB/ARA-12/01; 2012.
3. National Transportation Safety Board. Review of U.S. civil aviation accidents, 2007–2009. Washington, D.C.: National Transportation Safety Board; Report No.: Annual Review NTSB/ARA-11/01.
4. Boschen AC, Jones RK. Aviation language problem: Improving pilot-controller communication. *Proceedings of 2004 international professional communication conference*; 2004 Sep 29–Oct 2; Min neapolis, USA. Piscataway: IEEE Press; 2004. p.291–9.
5. Billings CE, Cheaney ED. Information transfer problems in the aviation system. Washington, D.C.: NASA; 1981. Report No.: NASA-TP-1875.
6. Schroeder D, Bailey L, Pounds J, Manning C. A human factors review of the operational error literature. Oklahoma City, OK: FAA Civil Aerospace Medical Institute; 2006. Report No.: DOT/FAA/AM-06/21.
7. Bordes A, Glorot X, Weston J, Bengio Y. A semantic matching energy function for learning with multi-relational data-Application to word-sense disambiguation. *Mach Learn* 2014;**94**(2):233–59.
8. Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the national conference on artificial intelligence*; 2006 Jul 16–20; Boston, USA. Palo Alto: American Association for Artificial Intelligence; 2006. p.775–80.
9. Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans Knowl Discov Data* 2008;**2**(2):1–25.
10. Socher R, Huang EH, Pennington J, Ng AY, Manning CD. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Proceedings of annual conference on neural information processing systems*; 2011 Dec 12–14; Granada, Spain. Red Hook: Curran Associates Inc.; 2011. p.801–9.

11. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [Internet]. 2013 [cited 2013 Sep 7]. Available from: http://cn.arxiv.org/pdf/1301.3781v3.
12. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems 26*; 2013 Dec 5–10; Lake Tahoe, USA. La Jolla: Neural Information Processing Systems Foundation; 2013. p.3111–9.
13. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. *Proceedings of 2014 conference on empirical methods in natural language processing*; 2014 Oct 25–29; Doha, Qatar. Stroudsburg: Association for Computational Linguistics (ACL); 2014. p. 1532–43.
14. Collobert R, Weston JA unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on machine learning*; 2008 Jul 5–9; Helsinki, Finland. New York: Association for Computing Machinery (ACM); 2008. p.160–7.
15. Meek C. Semantic parsing for single-relation question answering. *Proceedings of 52nd annual meeting of the association for computational linguistics*; 2014 Jun 22–27; Baltimore, USA. Stroudsburg: Association for Computational Linguistics (ACL); 2014. p.643–8.
16. Li ZC, Tang JH. Weakly supervised deep matrix factorization for social image understanding. *IEEE Trans Image Process* 2017;**26**(1):276–88.
17. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modeling sentences [Internet]. 2014 [cited 2014 Apr 8]. Available from: http://cn.arxiv.org/pdf/1404.2188v1.
18. Hu BT, Lu ZD, Li H, Chen QC. Convolutional neural network architectures for matching natural language sentence. *Advances in neural information processing systems 27*; 2014 Dec 8–13; Montreal, Canada. La Jolla: Neural Information Processing Systems Foundation; 2014. p.2042–50.
19. Socher R, Pennington J, Huang EH, Ng AY, Manning CD. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proceedings of conference on empirical methods in natural language processing*; 2011 Jul 27–31; Edinburgh, UK. Stroudsburg: Association for Computational Linguistics; 2011. p.151–61.
20. Mikolov T, Zweig G. Context dependent recurrent neural network language model. *Proceedings of IEEE workshop on spoken language technology*; 2012 Dec 2-5; Miami, USA. Piscataway: IEEE Press; 2013.p.234–39.
21. Kenter T, Rijke MD. Short text similarity with word embeddings. *Proceedings of the 24th ACM international conference on information and knowledge management*; 2015 Oct 19–23; Melbourne, Australia. New York: Association for Computing Machinery; 2015. p.1411–20.
22. Lai SW, Liu K, He S, Zhao J. How to generate a good word embedding. *IEEE Intell Syst* 2016;**31**(6):5–14.
23. Li Z, Liu J, Tang J, Lu H. Robust structured subspace learning for data representation. *IEEE Trans Pattern Anal Mach Intell* 2015;**37**(10):2085–98.
24. Liu B, Huang M. A sentence interaction network for modeling dependence between sentences. *Proceedings of the 54th annual meeting of the association for computational linguistics*; 2016 Aug 7–12; Berlin, Germany. Stroudsburg: Association for Computational Linguistics; 2016. p.558–67.
25. Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *15th annual conference of the international speech communication association*; 2014 Sep 14–18; Singapore. Baixas, France: ISCA; 2014. p.338–42.
26. Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. *Proceedings ofthe 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the asian federation of natural language processing*; 2015 Jul 26–31; Beijing, China. Stroudsburg: Association for Computational Linguistics (ACL); 2015. p.1556–66.
27. He H, Wieting J, Gimpel K, Rao J. UMD-TTIC-UW at SemEval-2016 task 1: Attention-based multi-perspective convolutional neural networks for textual similarity measurement. *Proceedings of 10th international workshop on semantic evaluation*; 2016 Jun 16–17; San Diego, USA. Stroudsburg: Association for Computational Linguistics (ACL); 2016. p.1103–8.
28. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on machine learning*; 2008 Jul 5-9; Helsinki, Finland. New York: Association for Computing Machinery (ACM); 2008. p.160-7.
29. Mikolov T, Karafiátet M, Burget L. Recurrent neural network based language model. *Proceedings of the 11th annual conference of the international speech communication association*; 2010 Sep 26–30; Makuhari, Japan. Lous Tourils: International Speech Communication Association; 2010. p.1045–8.
30. Mikolov T, Kombrink S, Burget L. Extensions of recurrent neural network language model. *Proceedings of IEEE international conference on acoustics, speech, and signal processing*; 2011 May 22–27; Prague, Czech Republic. Piscataway: IEEE Press; 2011. p.5528–31.
31. Sundermeyer M, Ney H, Schlüter R. From feedforward to recurrent LSTM neural networks for language modeling. *IEEE ACM Trans Audio Speech Lang Process* 2015;**23**(3):517–29.
32. Gers FA, Schmidhuber J. Recurrent nets that time and count. *Proceedings of the international joint conference on neural networks*; 2000 Jul 24–27; Como, Italy. Piscataway: IEEE Press; 2000. p.189–94.
33. Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with LSTM recurrent networks. *J Mach Learn Res* 2003;**3**(1):115–43.
34. Jia GM, Lu YJ, Lu WB, Shi YH, Yang JF. Verification method for Chinese aviation radiotelephony readbacks based on LSTM-RNN. *Electron Lett* 2017;**53**(6):401–3.
35. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. *Proceedings of the 30th international conference on machine learning*; 2013 Jun 16–21; Atlanta, USA. New York: Association for Computing Machinery (ACM); 2013. p.1139–47.
36. Lungu R, Lungu M. Automatic landing system using neural networks and radio-technical subsystems. *Chin J Aeronaut* 2017;**30**(1):399–411.