Mechatronic Systems and Control, 2022
An Open Access Paper

LSTM-ATTENTION TEXT CLASSIFICATION METHOD COMBINED WITH KEY INFORMATION

Jinbao Yang,* Min Ma,** Yu Fu,* and Yanhong Gu***

Abstract

Aiming at the problem that the text information could not be fully understood in traditional text classification, this paper proposed a recurrent neural network classification model based on the attention mechanism combined with key information and gave a sentence points extraction method. The semantics of the text was first expressed by the sentence points, which could fully express the semantic information and enrich its feature, so as to solve the problem of the feature sparsity in the text classification. Then, the bidirectional long short-term memory network with the attention mechanism was selected as the classifier for learning. Set the word vector and sentence vector as the input of the network, respectively, the outcomes could be spliced to obtain the final category. Experimental results show that this method could improve the accuracy of text classification.

Key Words

Text classification, points of text, attention mechanism, bidirectional long short-term memory (LSTM) network

1. Introduction

In recent years, machine learning and neural network have been widely used in target recognition [1], fault diagnosis [2] and motor control [3]. And with the rapid formation of the Internet industry and the widespread application of computer networks, a large number of text data have been imported into the network, showing an explosive growth trend [4]. Effective classification of these data is the premise of reasonable management and use of them. Therefore, many scholars have conducted research on text classification, such as Naive Bayesian (NB), support vector machine (SVM), decision tree, neural network learning and others. Literature [5]–[9] improved the traditional machine-learning and deep-learning methods, and applied them to the research of text classification, which achieved remarkable results. At present, the improved long short-term memory attention (LSTM-Attention) model based on recurrent neural networks has been widely used in the field of deep learning. Literature [10] improved the LSTM network by introducing the attention mechanism and applied it to news text classification. Besides, [11] and [12] used the LSTM-Attention model to study the semantic information, which greatly improved the accuracy of keyword extraction in text. Literature [13] put the word vector into bidirectional LSTM for feature extraction and finally emphasized the points according to the weight of the attention mechanism.

However, the LSTM-Attention model used above is only based on word vectors and cannot describe the complex semantic information of text effectively. Literature [14] expressed sentences in the text by combining words and syntactic structure which relies heavily on the analysis of syntactic. To solve this problem, document [15]–[17] studied the processing of text at sentence, and it makes the characteristics of text can be expressed by the sentence points. Furthermore, [18] proposed a method for generating text paragraph vectors (PV-DM). It used paragraph vectors to classify texts, which can map sentences of any length into vectors of a certain length. This method helped to analyse the text at the sentence level and improved the accuracy of text operations. Unfortunately, the components in sentences are normally redundant and complicated. Directly using sentence information to describe the text often leads to information redundancy. That is, while the characteristics related to the theme of the text are emphasized, the importance of some irrelevant information is also increased.

To enhance the main features of the text reasonably is an issue of great concern in the field of text classification. To this end, we proposed a recurrent neural network model of text classification based on the attention mechanism combined with key information and gave the method of sentence points extraction. This method can not only comprehensively represent the semantic information of the

* Department of Information Security, Naval University of Engineering, Wuhan 430033, China; e-mail: 252405059@qq.com, fuyu0219@163.com
** Naval Petty Officer Academy, Bengbu 233012, China; e-mail: hxmamin@163.com
*** Advanced Manufacturing Engineering Institute, Hefei University, Hefei 230601, China; e-mail: guyh@hfuu.edu.cn
Corresponding author: Yu Fu
Recommended by Jinbao Yang
(DOI: 10.2316/J.2022.201-0331)

text to solve the problem of feature sparsity in traditional text classification but also express the semantics of the text through sentence points to enrich the feature semantic information. Meanwhile, it used the LSTM network as a classifier and combined with the attention mechanism to adjust the weight of network output, which can improve the accuracy of text classification.

2. Long Short-Term Memory-Attention Model Combined with Key Information

To make full use of the semantic information in the sentences, the paper used the key points of sentence information to improve the traditional LSTM-Attention model and integrated the point events into the model. Through training, the model can automatically obtain the relationship between the key information of word, sentence and the text category and finally get the splicing method of the output of the word vector and the sentence point vector, so as to achieve accurate text classification.

2.1 Model of Long Short-Term Memory-Attention

The LSTM model based on the attention mechanism refers to introducing the attention mechanism [19] into the LSTM model [20]. It is different from the traditional LSTM network because it introduces attention into output of the hidden layer, that is, the weight, which can be expressed as:

$$\begin{aligned} H &= \sum_{i=1}^{n} \alpha_i h_i \\ \alpha_i &= \frac{e^{V \cdot \tanh(h_i W + b)}}{\sum_{k=1}^{n} e^{V \cdot \tanh(h_k W + b)}} \end{aligned} \tag{1}$$

where V and W are the parameters in the attention network model, b is the offset, $V \cdot \tanh(h_i + b)$ is the score of the hidden state at time i and α_i is the weight obtained by normalizing the attention through the *softmax* function at all times [21]. In the calculation process, the network is deep, and the number of iterations is large, which means overfitting is prone to occur. Therefore, when using LSTM-Attention for text classification operations, the Dropout method is often used to prevent overfitting.

2.2 Construction of Long Short-Term Memory-Attention Model Combined with Sentence Key Points

On the basis of the LSTM-Attention model, the paper added a connection associated with the sentence key points. The basic structure of the new LSTM-Attention model based on key information is shown in Fig. 1.

The specific process can be described as follows:

(1) Segment the word and tag the part-of-speech of the text to obtain the word vectors $X = (x_1, x_2, \cdots, x_m)$. Then extract sentence key information according to the word segmentation results and generate sentence key vectors using the PV-DM method $Core = (c_1, c_2, \cdots, c_m)$;

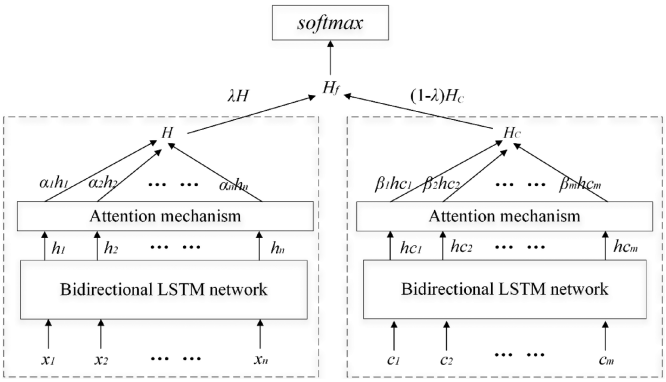

Figure 1. The model of LSTM-Attention combined with key information.

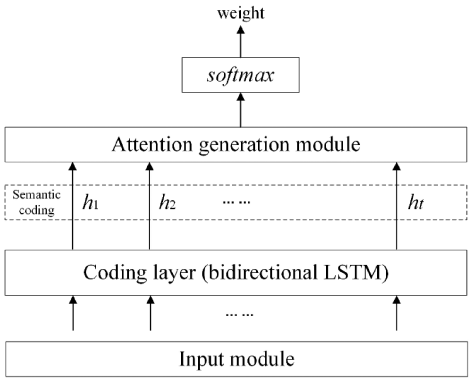

Figure 2. Attention generation model.

(2) Construct the LSTM-Attention network for word vectors and sentence point vectors, respectively, where the network parameters are obtained through learning and training. It is noteworthy that the LSTM served as the encoder in the attention generation model has nothing to do with the LSTM classifier in the LSTM-Attention model that introduces key information. The attention generation model in this paper is represented in Fig. 2. Assuming the input is $S = (s_1, s_2, \cdots, s_t)$, the semantic code obtained after the coding layer can be expressed as: $h_t = f(U \cdot s_t + W \cdot h_{t-1} + b)$, combined with the calculation of the LSTM network, the weight can be obtained as:

$$\begin{aligned} \alpha = softmax(&\tanh(h_t \cdot W_i + b_i) \\ &+ \tanh(h_{t-1} \cdot W_c + b_c) \cdot W_\alpha + b_\alpha) \end{aligned} \tag{2}$$

Using the above attention generation model, one can obtain the attention weights of the words and sentences, respectively, and then get the weighted output H and H_C.

(3) Splice the calculation results that obtained by the word vector and the sentence point vector, calculate the final result according to the proportion of the word vector and sentence vector calculation results and obtain the

进到 **银行** 现存的症结，人们的第一反应往往是不良 **贷款比率** 偏高、风险评估实战经验不足、**消费** 性 **金融产品** 缺失、企业管理标准不够完善等等。但另一方面，不可忽视的是，国内 **银行** 的品牌 **建设** 也存在某种滞后—鲜有差异化的品牌定位、品牌 **经营** 思维和以 **客户** 为本的鲜明形象，**顾客** 感受到的环境和 **服务** 面目雷同，甚至千脆一模一样。

Figure 3. Parts that directly relate to the topic in the sentence (Chinese text).

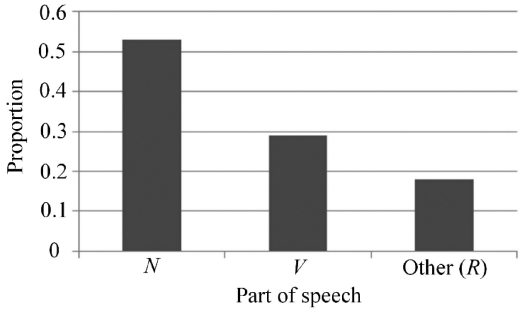

Figure 4. Distribution statistics of part-of-speech.

probability of belonging to the class:

$$H_f = \lambda \cdot H + (1 - \lambda) \cdot Hc$$
$$p = softmax(H_f) \tag{3}$$

where λ is a balance factor.

3. Sentence Key Information Extraction Algorithm

Sentence key information refers to the components that can reflect the semantics of the main body of the sentence. It is the condensation of sentence without changing the semantic information of the original sentence. Normally, the sentences of normative text are relatively long, the context cohesion is compact, the structure is complete and its grammar is standardized. Components used as connecting links often appear in the sentences, such as "in general" and "some experts think". The small correlation with the topic of the text or sentence will bring redundancy to the text analysis, which will make a greater cost. Therefore, such situations should be avoided in the process of text manipulation. As illustrated in Fig. 3, only the parts marked in different colour have a greater direct relationship to the text topic, and the other parts have no direct connection to the type of text. If the information is also taken into account when classifying the text, it will lead to redundancy of information, which has a certain negative impact on the results of text classification.

When extracting the key points of a sentence, the first place is to distinguish which component belongs to the main key information. In this paper, the main judgment basis is the part-of-speech of the words in the sentence. And the final retention information was determined according to the statistics of the correlation between the features in the experimental data set and the text classification. This paper used the chi-square statistic method. When the chi-square statistic of the feature to the text category is greater than 50, its part-of-speech distribution is shown in Fig. 4. It can be found that most of the words related to the text category are nouns or verbs.

The elements of the sets N, V and R are all part-of-speech, which can be expressed as [22]:

$$\begin{aligned} N &= \{n, nr, nr1, nr2, nrj, nrf, ns, nsf, nt, nz, nl, ng\} \\ V &= \{v, vd, vn, vshi, vyou, vf, vx, vi, vl, vg\} \\ R &= \{r, rr, rz, rzt, rzs, rzv, ry, ryt, rys, ryv, rg\} \end{aligned} \tag{4}$$

Therefore, the specific process of sentence key information extraction can be described by the following pseudocode algorithm:

Input: *text*
Output: *point*
1. *Sentence* ← *text segmented*
2. *m* ← *the num of sentence*
3. (*word*, *pos*) ← *sentence segmented*
4. *for i from* 1 *to m*
5. *n* ← *the num of word in sentence*[*i*]
6. *for j from* 1 *to n*
7. *if pos*[*j*]∈ $N \cup V \cup R$
8. *point*[*i*] ← *word*[*j*]
9. *end if*
10. *end for*
11. *end for*
12. *return point*

To make the main components of the sentence as less as possible, the pronoun part was also kept in the extraction in this paper. The sentence components were filtered according to the part-of-speech of the words in the sentence, and only the parts that could represent the core points of the sentence are left. It could prevent the redundancy in the sentences without changing the semantic information of the original sentence and improve the subsequent operability.

4. Simulation and Analysis of Results

4.1 Simulation

To verify the feasibility of the LSTM-Attention classification method introduced in this paper, the simulation experiment was conducted, and the specific configuration was as follows:

Experimental environment: Anaconda3.7, Keras, Jieba, Gensim.

Experimental indicators: The analysis indicators for natural language processing mainly include *precision*, *recall* and *F1-score*. Among them, the *precision* is the ratio of the number of the correctly identified items to the total identified items, the *recall* indicates how much of all items that should have been found were found and the *F1-score* is the weighted harmonic average of the *precision* and *recall* [23].

Experimental data: The experimental data used in this paper came from the Sogou corpus, that is, Sogou News data (Sogou CS), which collects news data from Sohu News that covers 18 channels including China, international,

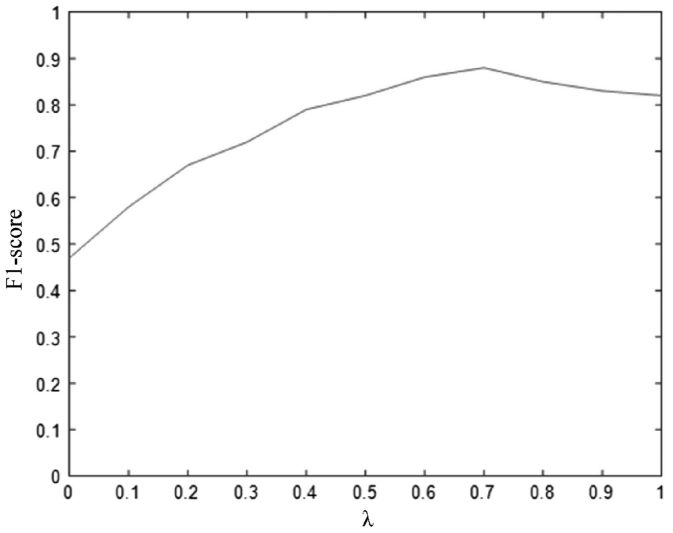

Figure 5. F1-score with different values of λ.

social, entertainment, *etc.* And the data package contained the data including URL, title, text content, *etc.*, which could be downloaded from the Sogou laboratory in ". Dat" format, with a size of 1.43 GB.

In the experiment, the parameters of the LSTM-Attention model were set as follows: the number of word vector dimension was 100, the number of key vector dimension was 100, the number of LSTM hidden layer units in Attention was 132, the number of LSTM hidden units in the coding layer was 132, the number of the pooling layer was 2, the initial learning rate was 0.001, the momentum was 0.95 and the Dropout was set to 0.4.

In addition, it needed to determine λ, the value of the balance factor in the above model. In the corpus, it took 1,000 pieces of news from the four channels of economy, technology, sports and entertainment, of which 800 pieces were for training sessions and 200 pieces for tests. It recorded the F1-score of the experiment under different balance factors. The results are displayed in Fig. 5.

According to the experimental results, it could be found that in the case of $\lambda \leq 0.7$, the F1-score gradually increased with the increase of λ, indicating that the word vector was very important for the text classification. While $\lambda \geq 0.7$, the F1-score gradually decreased, which showed that the continuous increase of the word vector component has a negative effect on the classification. When considering word vectors, some links in the text that related to the text category would be ignored, and it could be found that word vector's influence on text classification is higher than that of sentence vector. Observation showed that the F1-score of the experiment achieved maximum, when the value of λ was 0.7.

When it conducted text classification experiment on the rest 800 pieces of news, the results could be shown in Table 1.

4.2 Analysis of Result

In this section, comparative experiments were conducted to verify the feasibility of the purposed method. Except for the input vectors and classifiers, other parameters of the model were consistent with those in the previous experiment. The details are listed in Table 2.

The results, including the *Precision*, *Recall* and *F1-score*, are shown in Table 3. As we can see, the LSTM-Attention model based on the sentence vector had the worst experimental effect, which was because that method only took key points of sentences as input. Although the semantic information of the text was considered, only a small part of the subject information was retained in the extraction process when sentence vector was used. It would lead to the neglect of some useful information, which could result in unreliable experimental results. The LSTM-Attention model combined with key information had improved experimental results compared to the LSTM model, which indicated that the introduction of the attention mecha-

Table 1
The Results of LSTM-Attention Text Classification Combined with Key Information

Type	Financial	Scientific	Sports	Entertainment	Total Number
Financial	181	4	6	9	200
Scientific	6	188	4	2	200
Sports	5	3	185	7	200
Entertainment	8	4	9	179	200

Note: rows indicate the actual type of news, and columns indicate the experimental results.

Table 2
The Settings of Contrast Experiments

Experiment	Input	Classifiers
LSTM-Attention based on word vector	Word vector	LSTM-Attention
LSTM-Attention based on sentence vector	Point vector	LSTM-Attention
LSTM with key points	Word vector and point vector	LSTM
LSTM-Attention with key points	Word vector and point vector	LSTM-Attention

Table 3
Comparison of Experimental Results

Method	*Precision*	*Recall*	*F1-Score*
LSTM-Attention based on word vector	0.8215	0.8200	0.8203
LSTM-Attention based on sentence vector	0.4735	0.4638	0.4668
LSTM with key points	0.8726	0.8725	0.8724
LSTM-Attention with key points	0.9163	0.9163	0.9162

Table 4
Results of Different Classification Method

Method	NB	SVM	CNN	LSTM-Attention with key points
Financial	0.897	0.898	0.918	0.905
Scientific	0.940	0.9233	0.960	0.942
Sports	0.882	0.904	0.910	0.916
Entertainment	0.864	0.870	0.9007	0.902

Table 5
Results of the Experiments

Method / Type	*Precision*		*Recall*		*F1-value*	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Financial	0.9050	0.8939	0.9050	0.8850	0.9050	0.8894
Scientific	0.9447	0.9343	0.9400	0.9250	0.9424	0.9296
Sports	0.9069	0.8932	0.9250	0.9200	0.9158	0.9064
Entertainment	0.9086	0.8939	0.8950	0.8850	0.9018	0.8894

nism could help to distinguish the importance of output at different times. Compared with the LSTM-Attention model based on the word vectors, the proposed method improved in all three indicators. It could illustrate that the key information could effectively highlight the main components in the text, so as to improve the effectiveness of text classification.

To further prove the advantages of the text classification method that introduces key information, NB, SVM and convolutional neural network (CNN) were selected as the experimental comparisons. The F1-scores of the results are shown in Table 4.

It could be found that CNN and the LSTM-Attention with key information classification method were significantly better than the two methods of NB and SVM. That was mainly because when the sample data reached a certain level, the deep learning method could rely on its strong learning ability to mine the features in the text comprehensively. Except for the F1-score in scientific texts was slightly lower than CNN, the F1-score of LSTM-Attention classification method combined with key information in the other three types was higher than that of CNN method. In general, the average F1-score of LSTM-Attention method combined with key information was higher than several other commonly used text classification methods.

Another experiment was conducted to compare the method of extracting sentence key points with the method of using the whole sentence. In the method 1 (LSTM-Attention with key information), it needed to simplify the sentence before obtaining the corresponding key points vector, while the method 2 used PV-DM to directly convert the sentence into a sentence vector. The classification results of different types of news are shown in Table 5. The analysis showed that the method of introducing key information is better than the method without simplification in *precision*, *recall* and *F1-score*, which indicate that the method can effectively remove the redundancy in the sentence and increase the proportion of sentence's main components (semantics) in the classification. Although the LSTM-Attention method combined with key information might have removed some words which could be related to categories when constructing sentence vectors, the word vectors used in the classification process could compensate for the loss of components caused by sentence simplification to a certain extent.

5. Conclusion

In the traditional text classification process, word features cannot fully represent the text information, and sentence feature pairs will lead to redundancy. This paper proposed a LSTM-Attention classification method which was based on attention mechanism and combined key information. This method expressed the semantics of the text through the key information of the sentence, which increases the influence of the semantics of the text subject in the classification to a certain extent. Then it used the LSTM network as the learning classifier and adjusted the network output weight through the attention model, so that the accuracy of text classification could be effectively improved. However, when extracting the main points of the text, this method sometimes ignored the useful modification information in the sentence, which might affect the accuracy of text classification. This will be the next work to be improved in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61672531, in part by the Natural Science Foundation of Anhui Province under Grant 2008085QF316.

References

[1] W. Liu, Q. Wang, and Q. Guo, Automatic radar waveform recognition based on neural network, *Mechatronic Systems and Control*, *46*(2), 2018, 92–96.

[2] Q. Wang and X. Wang, A fault detection diagnosis predict observer based on resource allocation network, *Mechatronic Systems and Control*, *50*(2), 2022, 96–101.

[3] J. Tavoosi, A novel recurrent type-2 fuzzy neural network for stepper motor control, *Mechatronic Systems and Control*, *49*(1), 2021, https://doi.org/10.2316/J.2021.201-0097.

[4] W. Hong, W. Wang, Y. Weng, *et al.*, Stock price movements prediction with textual information, *Mechatronic Systems and Control*, *46*(3), 2018, 141–149.

[5] X.M. Song, Research on Chinese information classification based on improved Bayesian algorithm, *Beijing University of Posts and Telecommunications*, 2019.

[6] J.M. Cui, J.M. Liu, and Z.Y. Liao, Study on text classification technology based on SVM algorithm, *Computer Simulation*, *30*(02), 2013, 299–302.

[7] F. Lei, Research on text classification based on neural network and decision tree and its application, *University of Electronic Science and Technology of China*, 2018.

[8] K. Xiao, Z. Zhang, and J. Wu, Chinese text sentiment analysis based on improved Convolutional Neural Networks, *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (IEEE, 2016), 922–926.

[9] J. Wang and Z. Cao, Chinese text sentiment analysis using LSTM network based on L2 and Nadam, *2017 IEEE 17th International Conference on Communication Technology (ICCT)* (IEEE, 2017), 1891–1895.

[10] W.F. Lan, W. Xu, D.Z. Wang, *et al.*, Classification of Chinese news text based on LSTM-Attention, *Journal of South-Central University for Nationalities (Natural Science Edition)*, *37*(03), 2018, 129–133.

[11] Q.L. Zhao, X.D. Cai, B. Li, *et al.*, Text feature extraction method based on LSTM-Attention neural network, *Modern Electronic Technique*, *41*(08), 2018, 167–170.

[12] Y.P. Wang, M.X. Song, W. Wang, *et al.*, Analysis of the tendency of decision results based on deep learning, *Application Research of Computers*, *36*(02), 2019, 335–338.

[13] S.W. Tian, W. Hu, L. Qi, E. Tu, B.L.Y. Yi, J.G. Zhao, and W. Li, The time series relationship identification of Bi-LSTM Uyghur events combined with Attention Mechanism, *Journal of Southeast University (Natural Science Edition)*, *48*(03), 2018, 393–399.

[14] R. Socher, C.C. Lin, A.Y. Ng, *et al.*, Parsing natural scenes and natural language with re-cursive neural networks, *ICML 2011: Proceedings of the 28th International Conference on Machine Learning* (OMNI Press, Bellevue, Washington, 2011), 129–136.

[15] Y. Bengio, R. Ducharme, P. Vincent, *et al.*, Neural probabilistic language model, *Journal of Machine Learning Research*, *3*(6), 2003, 1137–1155.

[16] T. Mikolvol, I. Sutskever, and K. Chen, Distributed representations of words and phrases and their compositionality, *Proceedings of the 26th International Conference on Neural Information Processing System* (Lake Tahoe: Curran Associates Inc., 2013), 91–100.

[17] T. Mikolov, K. Chen, G. Corrado, *et al.*, Efficient estimation of word representations in vector space [EB/OL]. http://arxiv.org/pdf/1301.3781.pdf (accessed June 20, 2017/ September 01, 2019).

[18] Q. Le and T. Mikolov, Distributed representations of sentences and documents, *Proceeding of the 13th Conference on International Conference on Machine Learning* (Cambridge, MA: MIT Press, 2013), 1188–1196.

[19] G. Liu and J. Guo, Bidirectional LSTM with attention mechanism and convolutional layer for text classification, *Neurocomputing*, *337*, 2019, 325–338.

[20] B. Keivan and G. Reza, Hierarchical LSTM network for text classification, *SN Applied Sciences*, *1*(9), 2019, 1124.

[21] G. Bouchard, Efficient bounds for the softmax function and applications to approximate inference in hybrid models, *NIPS workshop on approximate inference in hybrid models*, 2007.

[22] J.Y. Sun, Jieba Chinese word segmentation [EB/OL]. http://pypi.python.org/pypi/jieba/ (accessed June 20, 2017/September 01, 2019).

[23] L. Derczynski, *Complementarity, F-score, and NLP Evaluation* (Portorož, Slovenia: European Language Resources Association (ELRA), 2016).

Biographies

Jinbao Yang received the B.E. degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, and received the M.S degree in Communication and Information System, Naval University of Engineering. He is currently an associate professor with the department of information security, Naval University of Engineering, Wuhan, China. His research interests include information security and intelligence security.

Yu Fu received the B.E. degree in automation from Wuhan University of Technology, Wuhan, China, in 2019. He is currently pursuing the M.S. degree in the School of Automation, Wuhan University of Technology, Wuhan, China. His research interests include 3D shape measurement and camera calibration.

Min Ma graduated from Nanjing Electronic Engineering Academy, major in information engineering, in Nanjing, Jiangsu Province. He is currently a network engineer with Naval Petty Officer Academy, Bengbu, Anhui, China. His research interests include information security and intelligence security.

Yanhong Gu received a Ph.D. degree in optics from the University of Science and Technology of China, Hefei, China, in 2017. She is currently a lecturer with the Advanced Manufacturing Engineering Institute, Hefei University, Hefei, China. Her research interests include photoelectric detection and processing.