EasyChair Preprint

№ 5236

A Hybrid and Regenerative Model Chat Robot Based on LSTM and Attention Model

Dongyang Gao, Junwu Zhu and Fudong Li

March 30, 2021

A hybrid and regenerative model chat robot based on LSTM and Attention Model

Dongyang Gao[a] , Junwu Zhu[a], Fudong Li[a]
[a]School of Information Engineering, Yangzhou University, Yangzhou Jiangsu, China

ABSTRACT

Aiming at the situation that retrieval chat robot relies too much on predefined responses and the training requirements of generative chat robot are too high, a hybrid and regenerative model text chat robot based on LSTM and Attention-model is designed. Due to the retrieval model can only handle scenarios with predefined responses, and a generative model with strong learning ability will produce grammatical errors in certain scenarios. Therefore, firstly,doing text processing based on corpus, and then the retrieval model generates a candidate data set, and the candidate data set is trained by generating model to obtain the final model. The experimental comparison results show that the hybrid and regenerative model chat robot can effectively improve the model response quality compared to the single model chat robot, and accuracy improved by thirty percent.

Keywords: Deep learning,text generation,chat robot,NLP,seq2seq,attention model

1. INTRODUCTION

With in-depth research of artificial intelligence, deep learning technology represented by chatbot is very likely to become the entrance and interactive platform of Internet information services in the future. On the one hand, chatbot aims to realize the function of analyzing customer's questions through NLP and other technologies, and to generate responses through the database to find similar answers[1]. On the other hand, it is also an anthropomorphic dialogue model, a program that deep learning models communicate with humans. Compared with manual customer service, text interactive chatbot has three advantages: one is to optimize the user experience and reduce the complexity of solving customer problems; The second is to improve service efficiency, speed up the speed of response without limiting the service period, thereby saving service cost; The third is to facilitate the collection of user data and provide necessary assistance for product iterative optimization[2].

Compared with ordinary question-and-answer systems, almost all problems handled by chatbots that applied in specific fields are related to professions. Therefore, when building a suitable database and training objects, the response accuracy of text interactive chatbot will also be improved. At the same time, the increased performance of computer hardware such as CPU, GPU and the application of more efficient deep learning algorithms have also promoted the practical application of text interactive customer chatbot to a higher level. Although text interactive chatbots used in specific fields based on deep learning have achieved certain success, the traditional seq2seq model has many shortcomings, such as generating meaningless answers like "I don't know" and difficult to conduct multiple rounds of dialogue. In addition, limiting the length of text input also affects the performance of the model[3].

In response to the above problems, this paper uses a hybrid and regenerative chatbot based on LSTM and Attention model that combines retrieval model and generative model. The chatbot is completed from three aspects: system model construction, text processing, and improved seq2seq model. The experimental comparison results show that the hybrid and regenerative model chatbot can effectively improve the response quality compared to single model chatbot.

The chapters of this paper are arranged as follows:I will introduce some of the related work of this paper in chapter two, including the research status at home and abroad and the existing problems and solutions of hybrid model robots.The chapter three will be divided into two parts, the first part introduces the advantages and disadvantages of common models and the basic principles of the hybrid regenerative model proposed in the paper, the second part focuses on the more important expected processing process when training the model. The chapter four mainly introduces some theoretical basis of design in this paper:LSTM, NLP language model and attention model. The chapter five will focus on the experimental part of this paper: Experimental environment, experimental data processing and comparative analysis. Chapter six is the conclusion, which summarizes the overall design of this paper and focuses on the future research direction[4].

2. RELATED WORK

2.1 Research Actuality

In the process of the development and application of chatbot technology, retrieval models and generative models have been around for a long time, and the technology has become increasingly mature. The technology of fusing the above two models is the current mainstream research direction in the field of chat robot research.

Tan Menghua et al[5]. used the divergence-convergence clustering analysis method to design the user experience in the user's thinking to solve the problem of the robot's answer to the question and improve the accuracy of the reply when the robot and the user chat with each other.

Wu Shisong et al[6]. designed a dialogue generation mechanism based on seq2seq and attention model, using the maximum matching word segmentation algorithm to match semantics. According to the matching results, search for similar words in the dialogue and expand the dialogue keywords. This dialogue generation mechanism can better recognize sentences, improve the recognition rate of words, and ensure that the chatbot makes correct responses in actual conversations.

Xu Chang et al[7]. performed a supervised classification of the specific tasks to be performed by the chatbot, and then added the classification results to the chatbot model for supervised training. At the same time, a pre-training discriminator is added to optimize the seq2seq model. Under specific working conditions, it avoids the problem that the model response is not enough to meet the actual task requirements, effectively improves the quality of the model response, and improves the BELU score by 0.0116.

2.2 Hybrid chatbot model

Classical neural network models usually have gradient dissipation during text training[8]. This is because when training the text content, as the text interval between the current predicted position and related information continues to increase, the neural network may not be able to learn long-distance information[9]. Even in more complex language scenarios, the interval and length of useful information may limit the performance of recurrent neural networks. Therefore, in response to the above situation, this article introduces the attention mechanism and proposes a hybrid model chatbot based on LSTM and Attention-model[10]. The experimental comparison results show that the hybrid model chatbot can effectively improve the model response quality compared to a single model chatbot[11].

3. MODEL

3.1 Chatbot system model

The retrieval model is relatively easy to implement in structure, and it can directly use a database of predefined responses to select an appropriate output response according to the input and context[12][13]. However, the chat system based on the retrieval model will not generate any new text, that means the system's answer can only be related data that has been defined in the database[14-15].

The generative model does not rely too much on the predefined response of the database. The model generates a new response completely from scratch. Generative models are more based on machine translation technology, which can generate appropriate output data according to the context[16-17].

The hybrid and regenerative model has both a retrieval mode and a generation mode. The retrieval mode generates a candidate data set, takes the candidate data set as the input of the generative model, and obtains the final output through feature extraction and decoding of the candidate data set[18-19]. The schematic diagram of the hybrid and regenerative model is shown in Figure 1.

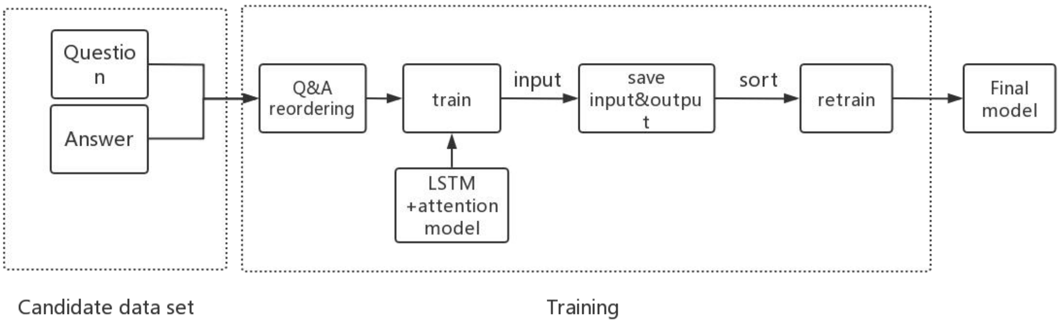

Figure 1. Schematic diagram of hybrid and regenerative model

3.2 Text processing

The flow chart of corpus processing is shown in Figure 2.

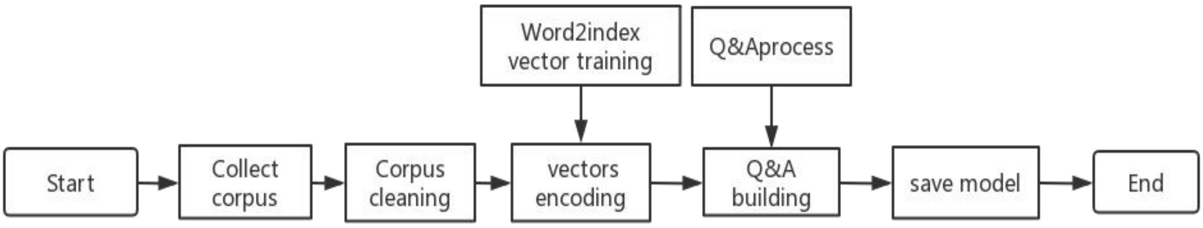

Figure 2. Flow chart of corpus processing

Corpus collection comes from the following scenario: chat records; movie dialogues; line fragments[20]. The content of corpus cleaning: extra spaces; irregular symbols; extra characters etc. Methods of corpus cleaning: regularization, segmentation, judgment of good and bad sentences. After the corpus is cleaned, it is just ordinary text, which is in the format of a string, and can't be directly trained, it can be used only after the encoding of the sentence vector[21-22].

Even the processed corpus is combined together, usually a TXT document, so it is necessary to split the question and answer pairs to separate the questions and answers. The corpus model preservation flowchart is shown in Figure 3.

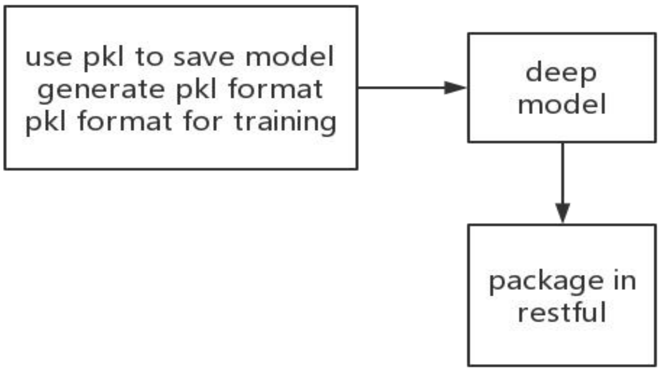

Figure 3. Flow chart of corpus model saving

4. THEORETICAL BASIS

4.1 LSTM

When the text interval between the current predicted position and related information is increasing, the simple cyclic neural network may not be able to learn long-distance information. Even in more complex language scenarios, the interval and length of useful information are both May limit the performance of recurrent neural networks[23-24].

Long short-term memory (LSTM) can effectively solve such problems. The performance of the recurrent neural network using the LSTM structure is better than that of the classic recurrent neural network[25]. Different from a single loop body structure, LSTM is a special network structure with three "gates", which can transmit information from the previous neural unit to the next neural unit[26]. The LSTM unit structure is shown in Figure 4.

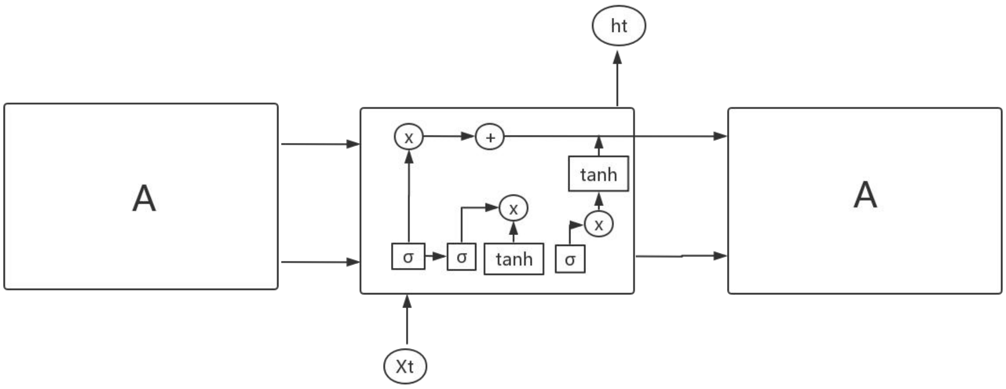

Figure 4. LSTM unit structure diagram

4.2 NLP language model

The main function of the language model is to calculate the probability that a word sequence constitutes a sentence, or to calculate the joint probability of a word sequence[27]. This can be used to judge whether the probability of a sentence is high or not, which does not conform to our expression habits[28].

Common models are: Uni-gram models (unary grammar statistical model), N-gram language model (N-gram model).

Uni-gram models：Find the probability of each word , and then multiply these probabilities[29].

$$\mathrm{p}(s) = p(w1) * p(w2) * .... * p(wn) \tag{1}$$

The condition for the establishment of this formula is that there is a hypothesis, that is, the condition-independent hypothesis, that every word is condition-independent[30].

For example: The weather today is very sunny.

P(now)=1/10, p(day)=1/10, until the period after it is 1/10.

N-gram model：The formula of the N-gram model is shown below

$$\mathrm{p}(w1, w2, ..., wt-1, wT) = \prod_{t=1}^{T} p(wt \mid w1, w2, ..., wt-1) \approx \prod_{t=1}^{T} p(wt \mid w_{t-n+1}^{t-1}) \tag{2}$$

The next word the model wants to predict[31].

$$\mathrm{w}_{\mathrm{t\text{-}n+1}}^{t-1} = w_{t-n+1}, w_{t-n+2,...,} \mathrm{w}_{t-1} . \tag{3}$$

Historical words in sentences. After that, use maximum likelihood estimation to optimize the model[32].

$$\sum_{t=1}^{T} \log p(w_t \mid w_{t-n+1}^{t=1}, \theta) \quad . \tag{4}$$

Count the number of occurrences of words in various situations, and then divide by a factor to normalize[33].

Although the N-Gram language model is highly interpretable and is easy to implement incrementally and parallel training, it still needs to solve the problem of data sparsity, such as the probability of words that have not appeared before being set to 0. At the same time, it is discrete Type variables, there is no way to measure the similarity between words[34-35].

4.3 Attention model

Although the Encoder in Seq2Seq can replace RNN with LSTM to enhance the information expression of the final semantic vector C on the long input sequence, because the traditional Seq2Seq model encodes the input sequence and the output semantic vector C is fixed, a vector cannot encodes all the information contained in the input sequence well, and the decoding stage is limited by the fixed-length vector representation, and the introduction of Attention model can effectively solve this limitation[36-37]. The schematic diagram of the Attention model is shown in Figure 5.

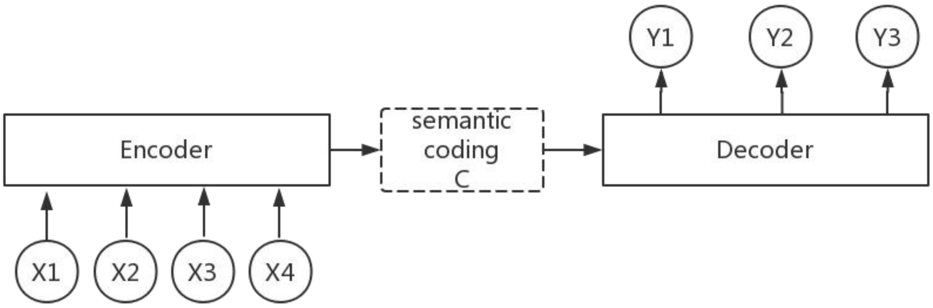

Figure 5. Schematic diagram of Attention Model

5. EXPERIMENT

5.1 Experimental environment and data

The experimental system is Windows 10, the GPU is 2070super, and the model uses the TensorFlow framework, developed by Google. The experimental data uses Renren's Xiaohuangji corpus, which contains about 450,000 corpus. After word segmentation is performed on the corpus, the dialogue content of the corpus is shown in Table 1.

Table 1. Examples of Corpus Q&A

Question	Answer
Do you know the result of the Red Sox game last night?	I closed it before I saw the end
What if there's an accident?	There will be no accident. You'll be safe.

5.2 Analysis of experimental results

In order to ensure the accuracy of the experimental comparison results, the basic parameters of the training model are kept consistent, and the parameter settings are shown in Table 2. n_epoch indicates the number of training rounds, batch_size indicates the number of samples selected for one training session, cell_type indicates the type of RNN neuron, and Attention_type indicates the type of attention mechanism used.

Table 2. Training parameters

Model	N_epoch	Batch_size	Cell_type	Attention_type
hybrid	15	128	LSTM	Bahdanau
generative	15	128	LSTM	None
retrieval	0	0	0	None

Based on the above experimental parameters, the training test was carried out, and the experimental results were shown in Figure 6.

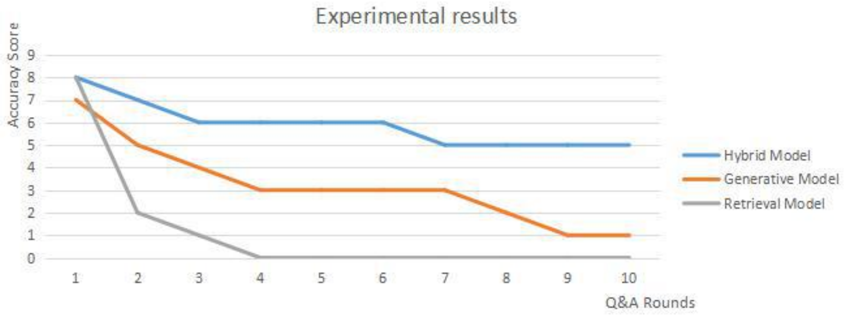

Figure 6. Experimental results

In addition, This paper also records the iteration of loss in the training process of generative model and hybrid and regenerative model, Figure 7 shows the comparison of loss during the training process, the blue line represents hybrid and regeneration model,the yellow line represents hybrid model. It can be seen from the figure that the loss of the hybrid and regenerative model, decreases rapidly and stably.

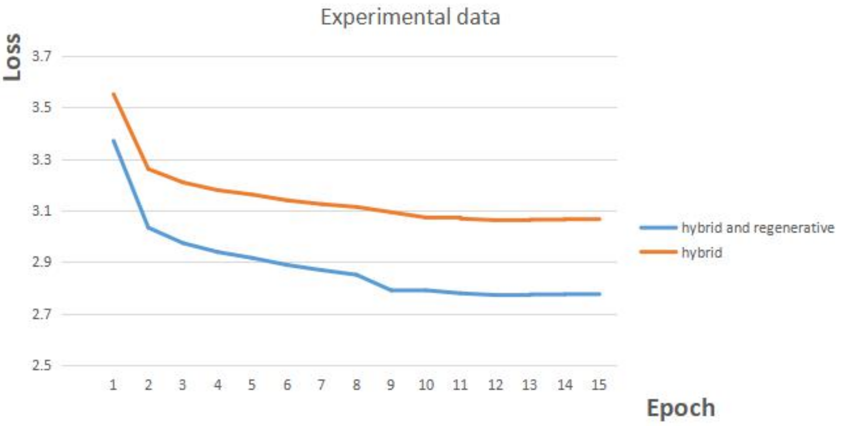

Figure 7. Comparison of loss during training

In the traditional generative model based on seq2seq, output of the model used be the most frequently used question and answer statements in the training process, for example: Who are you? I don't know; Hello;Ha ha etc, nonsense statement, to avoid such meaningless replies, We propose a hybrid and regenerative model: We shuffle the question and answer pairs of the existing database twice, and get two batches of data, One batch of data was used for initial model training, the second batch of data tested the model, save its input and output as the third batch of data, and this is the optimized Q&A training data, this can increase a certain degree of data relevance. And then the third batch of Q&A data is used to retrain the model, the final hybrid and regenerative model is obtained. Some experimental results are shown in the table below.

Table 3. Q&A test comparison

Question	Answer(hybrid)	Answer(hybrid and regenerative)
Who is Li Si?	who's?	Is a complete Dork
Hello?	Hello	Hello
Who am I?	I Ah	You are my master
Am I beautiful?	thank you	Very very very very very very good-looking
How's the weather today？	What's the weather like today	It's a fine day today.
Go out to play?	OK	Sure. What should I prepare

Analyzing the above experimental results, the following conclusions can be drawn:

First: In the case of multiple rounds of conversations that are more suitable for actual application scenarios, the retrieval model is almost incapable of making relevant responses based on the contextual text, while the generative model and the hybrid and regenerative model have a certain ability to respond.

Second: Whether it is a single round of dialogue or multiple rounds of dialogue, the hybrid and regenerative model which used LSTM and Attention-model has better model generalization ability than the generative model.

In summary, the hybrid and regenerative model chatbot based on LSTM and Attention-model designed this time can improve the quality of model responses more effectively than the generative model based on deep learning and the retrieval model based on database.

CONCLUSION

In view of the situation that retrieval chatbot relies too much on predefined responses and the training requirements of generative chatbot is too high, this paper proposes a hybrid and regenerative model text chatbot based on LSTM and Attention-model. This design can not only effectively reduce training costs, but also improve the training effect. It can also be applied to different professional fields by changing the corpus text. The experimental comparison results show that the hybrid and regenerative model chatbot can effectively improve the model response quality compared to the single model chatbot.

In future studies, the professional performance of chatbot can be further improved by optimizing the corpus text and perfecting the scoring mechanism, the generalization ability of the model can also be enhanced by optimizing the deep neural network to alleviate the gradient disappearance and overfitting in the model training process.

REFERENCES

[1] Zhang Liang. Research on chatbot in Combination of Retrieval and Generation[D].East China Normal University,2020.

[2] Guo Qixin,Yu Weihong,Li Chaohui.Design of Chat Robot System for Customer Service[J].Computer engineering & Software,2019,40(09):84-86.

[3] Wang Hao,Guo Bin,Hao Shaoyang,Zhang Qiuyun,Yu Zhiwen. Personalized dialogue content generation based on deep learning[J].Journal of Graphics,2020,41(02):210-216.

[4] Zhu Zeqi.Research on sensitive content recognition for chat robot[J]. INTELLIGENT COMPUTER AND APPLICATIONS,2020,10(03):218-222.

[5] Tan Menghua,Pan Xiaoyan. Research on Text Chatbot Conversational Reply Strategy[J]. Computer engineering & Software,2020,41(09):51-55.

[6] Wu Shisong,Lin Zhida.Research on dialogue generation mechanism of chat robot based on Seq2 seq and Attention model[J].AUTIMATION&INSTREMENTATIONAL,2020(07):186-189.

[7] Zhang Xin, Deng Zhuoheng, Jin Yifei, He Hongchen, Bai Ling. Design and Implementation of Chat Robot Based on NLP Seismic Science[J].Modern Information Technology,2020,4(11):77-79.

[8] Tom Young, Devamanyu Hazarika, Soujanya Poria, et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. CoRR, 2018, abs/1708.02709.

[9] Li Deng, Yang Liu. Deep Learning in Natural Language Processing[M]. Singapore: Springer, 2018.

[10] Ji Zongcheng, Lu Zhengdong, Li Hang. An Information Retrieval Approach to Short Text Conversation[J]. CoRR, 2014, abs/1408.6988.

[11] Wu Yu, Wu Wei, Xing Chen, et al. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(ACL’17), Vancouver, 2017.

[12] Ashish Vaswani,Noam Shazeer, Niki Parmar, et al. Attention Is All You Need[J]. CoRR, 2017, abs/1706.03762.

[13] Wu Weizhen. Research on Chat Robot Dialogue Based on seq2seq Model [D]. Nanjing University of Posts and Telecommunications,2019.

[14] Wang Qianming,Li Yin.Research on Personalized Chatbot Based on Deep Learning[J]. Computer Technology and Development,2020,30(04):79-84.

[15] Qi Jiayin,Hu Shuaibo,Zhang Ya. Application of Artificial Intelligence Chatterbot in Digital Marketing — Literature Review.

[16] Yang Ye. Research on Chatbot Based on Deep Learning[J]. Information Technology and Informatization, 2020(03):158-159.

[17] Yujie Liang,Liang Yujie,Yu Ying,Ouyang Wenhao. Intelligent chat robot in digital campus based on deep learning[J]. Journal of physics. Conference series,2020,1629(1).

[18] Lee Othelia EunKyoung,Davis Boyd. Adapting 'Sunshine,' A Socially Assistive Chat Robot for Older Adults with Cognitive Impairment: A Pilot Study.[J]. Journal of gerontological social work,2020.

[19] Yang Jingxian,Zhang Shuai,Xiang Yue,Liu Jichun,Liu Junyong,Han Xiaoyan,Teng Fei. LSTM auto-encoder based representative scenario generation method for hybrid hydro-PV power system[J]. IET Generation, Transmission & Distribution,2020,14(24).

[20] Wenjin Zhang,Jiacun Wang,Fangping Lan.Dynamic Hand Gesture Recognition Based on Short-Term Sampling Neural Networks[J].IEEE/CAA Journal of Automatica Sinica,2021,8(01):110-120.

[21] Where to Prune: Using LSTM to Guide Data-dependent Soft Pruning.[J]. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society,2020,PP.

[22] Information Technology - Data Extraction; Data on Data Extraction Reported by Researchers at Department of Computer Engineering (Can We Survive Without Labelled Data In Nlp? Transfer Learning for Open Information Extraction)[J]. Computer Technology Journal,2020.

[23] QIU XiPeng,SUN TianXiang,XU YiGe,SHAO YunFan,DAI Ning,HUANG XuanJing.Pre-trained models for natural language processing: A survey[J].Science China(Technological Sciences),2020,63(10):1872-1897.

[24] Johnny Torres,Carmen Vaca,Luis Terán,Cristina L. Abad. Seq2Seq models for recommending short text conversations[J]. Expert Systems With Applications,2020,150.

[25] Ábel Elekes,Adrian Englhardt,Martin Schäler,Klemens Böhm. Toward meaningful notions of similarity in NLP embedding models[J]. International Journal on Digital Libraries,2020,21(2).

[26] Hariharan Jayakumar,Madhav Sankar Krishnakumar,Vishal Veda Vyas Peddagopu,Rajeswari Sridhar. RNN based question answer generation and ranking for financial documents using financial NER[J]. S ā dhan ā,2020,45(1).

[27] Joffrey L. Leevy,Taghi M. Khoshgoftaar,Flavio Villanustre. Survey on RNN and CRF models for de-identification of medical free text[J]. Journal of Big Data,2020,7(1).

[28] Gaiping Sun,Chuanwen Jiang,Xu Wang,Xiu Yang. Short - term building load forecast based on a data - mining feature selection and LSTM - RNN method[J]. IEEJ Transactions on Electrical and Electronic Engineering,2020,15(7).

[29] Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach.[J]. IEEE journal of biomedical and health informatics,2020,PP.

[30] Ming Zhou,Nan Duan,Shujie Liu,Heung-Yeung Shum. Progress in Neural NLP: Modeling, Learning, and Reasoning[J]. Engineering,2020,6(3).

[31] Anas Almunif,Lingling Fan. Optimal PMU placement for modeling power grid observability with mathematical programming methods[J]. International Transactions on Electrical Energy Systems,2020,30(2).

[32] Wennian Yu,Il Yong Kim,Chris Mechefske. Analysis of different RNN autoencoder variants for time series classification and machine prognostics[J]. Mechanical Systems and Signal Processing,2021,149.

[33] Robotics; New Findings from Lanzhou University in the Area of Robotics Reported (Rnn for Solving Time-variant Generalized Sylvester Equation With Applications To Robots and Acoustic Source Localization)[J]. Computers Networks & Communications,2020.

[34] Lyan Verwimp,Hugo Van hamme,Patrick Wambacq. State gradients for analyzing memory in LSTM language models[J]. Computer Speech & Language,2020,61.

[35] Dian Yu,Shouqian Sun. A Systematic Exploration of Deep Neural Networks for EDA-Based Emotion Recognition[J]. Information,2020,11(4).

[36] Ghazanfar Ali,Myungho Lee,Jae - In Hwang. Automatic text - to - gesture rule generation for embodied conversational agents[J]. Computer Animation and Virtual Worlds,2020,31(4-5).

[37] Wei Fang,TianXiao Jiang,Ke Jiang,Feihong Zhang,Yewen Ding,Jack Sheng. A method of automatic text summarisation based on long short-term memory[J]. International Journal of Computational Science and Engineering,2020,22(1).