IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2025, vol. 3, no. 1, pp: 29–34

Chinese Language Model Adaptive Method Based on Recurrent Neural Network

Jiangjiang Li[1], Jiaxiang Wang[1], Lijuan Feng[1], and Yachao Zhang[1]

School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology
110064 Zhengzhou, China
Received Nov. 28, 2024; Revised and Accepted Feb. 20, 2025

Abstract. Deep learning is more and more widely used in natural language processing. Compared with the traditional n-gram statistical language model, Recurrent neural network (RNN) modeling technology has shown great advantages in language modeling, and has been gradually applied in speech recognition, machine translation and other fields. However, at present, the training of RNN language models is mostly offline. For different speech recognition tasks, there are language differences between training corpus and recognition tasks, which affects the recognition rate of speech recognition systems. While using RNN modeling technology to train the Chinese language model, an online RNN model self-adaption algorithm is proposed, which takes the preliminary recognition results of speech signals as corpus to continue training the model, so that the adaptive RNN model can get the maximum match with the recognition task. The experimental results show that the adaptive model effectively reduces the language difference between the language model and the recognition task, and the recognition rate of the system is further improved after the Chinese word confusion network is re-scored, which has been verified in the actual Chinese speech recognition system.

Keywords: Deep learning, Recurrent neural network, Chinese language model, Self-adaption algorithm.

1. Introduction

Speech recognition refers to the technology that machines can recognize and understand human speech signals into corresponding text or commands [1,2]. In speech recognition, continuous speech contains rich grammatical and syntactic information. The fundamental purpose of adding language models to the recognizer is to classify and model these grammatical and syntactic information, and find out the best word sequence to reduce the matching search range between speech feature vector sequence and word sequence. At this point, language models play an important role in the form of prior probabilities, incorporating various high-level non-acoustic knowledge into speech recognition systems [3-5].

The performance of the speech recognition system largely depends on the matching degree of the language model and the recognition task, and is strongly dependent on the environment. When the language model matches the recognition task topic, good recognition results can often be obtained; otherwise, the recognition performance deteriorates [6,7]. In practical applications, recognition tasks are often mixed with multiple and unpredictable topics, especially for telephone speech recognition, this feature is particularly obvious. The telephone voice is mostly spoken, which often involves multiple topics in content, and different speakers have different speaking styles. If there is sufficient spoken corpus, then the problem of matching the trained language model with the recognition task may be partially solved, but a large number of spoken corpus is not easy to collect. Therefore, how to quickly and accurately implement language model adaptation and match with the recognition task theme becomes a key problem [8-10].

The traditional language model adaptive technique [11] combines a general, well-trained language model and a domain-specific, inadequately trained model into a new model in some way. Therefore, this adaptive technology is often called topic adaptive or domain adaptive technology. There are two methods of combination: interpolation method and maximum entropy method. Interpolation method is commonly used, its biggest advantage is easy to implement, high computational efficiency, its disadvantage is difficult to ensure the integrity of the model, and it is difficult to achieve the best interpolation effect; The advantage of maximum entropy method is that it can achieve more optimal interpolation effect, but its disadvantage is that the calculation is large and the calculation efficiency is low. These methods have one thing in common, that is, they collect the adaptive corpus of the domain in advance when the domain is known, and then train it in an offline way [12,13]. If the domain changes, the model needs to be re-trained to determine the adaptive coefficient. However, sometimes the corpus obtained in advance is very small, and the field of use can only be determined when it is applied, especially for telephone speech recognition, the subject of the speaker cannot be predicted in advance, and the traditional language model adaptive technology will no longer be suitable [14,15].

In recent years, deep learning has been gradually applied to natural language understanding. Compared with the traditional n-gram statistical language model, recurrent neural network modeling technology [16,17] has shown great advantages in language model modeling and has been gradually applied in speech recognition, machine translation and other fields. In this paper, RNN modeling technology is first applied to the modeling of Chinese language model, and on this basis, an online adaptive algorithm based on RNN Chinese language model is proposed to carry out model adaptive adaptation on the preliminary recognition results of speech signals, that is, the results of speech recognition are used as training corpus after word segmentation and other processing and continue training on the basis of the original RNN model. By learning to update the weight matrix of the model, the model can better reflect the language distribution of the recognition task after self-adaptation, and then the recognition task is decoded again.

2. Speech Recognition System Framework

The basic block diagram of a complete speech recognition system is shown in Figure 1. Language model is an important part of the speech recognition system. The use of higher-level language knowledge can reduce the fuzziness of pattern matching on the basis of acoustic recognition, thus improving the accuracy of system recognition.

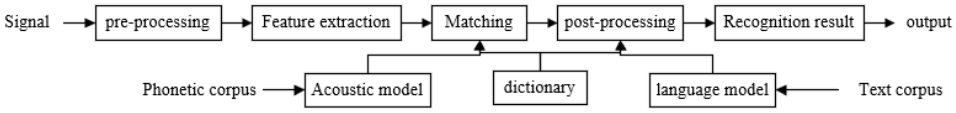

Fig. 1. Structure of Chinese speech recognition system

After preprocessing and feature extraction of the speech signal, the acoustic feature vector contained in the speech signal can be obtained, which is denoted as O. Natural language can be thought of as a random sequence, where every sentence or word in a text is a random variable with a certain distribution. Assuming that words (including single words in Chinese) are the smallest structural unit of a sentence, a reasonable meaningful statement S consists of a sequence of words $w = w_1, w_2, \cdots, w_N$. Starting from the Bayes principle, the process of speech recognition is to find out the word sequence with the greatest conditional probability under the current acoustic characteristics as the recognition result according to formula (1).

$$\hat{w} = arg\max_{w} P(w|o) = arg\max_{w} \frac{p(o|w)p(w)}{p(o)}. \tag{1}$$

$$p(w) = \prod_{i=1}^{N} p(w_i|w_1, \cdots, w_{i-1}). \tag{2}$$

The prior probability $p(o)$ of the signal waveform has nothing to do with the choice of word sequence w and cannot be calculated. $p(o|w)$ represents the possibility of output feature sequence based on a given sequence of words, which is modeled by acoustic model in speech recognition. And $p(w)$ represents the possibility of word sequence w, which is modeled by language model in speech recognition. Currently, the widely used n-gram language model holds that the probability of occurrence of each predictor is only related to the context of length $n-1$, that is,

$$P(w_i|w_1, w_2, \cdots, w_{i-1}) = P(w_i|w_{i-n+1}^{i-1}). \tag{3}$$

Usually the n value is 2 or 3, only the language information of the local context of the current word is considered, and its training often requires a large number of real training corpus. Compared with the traditional n-gram language model based on statistical rules, the RNN language model takes more historical information into account when predicting a word, so it can describe the long distance information in the sentence better [18-20].

2.1. Processing of Chinese Corpus

Before the training of Chinese text corpus, a series of processing must be carried out on the corpus. The processing flowchart of Chinese training corpus is shown in Figure 2. After a text corpus is obtained, it should be cleaned first, and the noisy information such as letters and punctuation marks in the coarse corpus should be deleted to remove

redundant information; There are often a large number of numbers in the corpus, and the normalization process mainly completes the normalization of linguistic numerals and converts Arabic numerals into corresponding Chinese characters. After the first two steps, there is only Chinese information in the corpus. Word segmentation is to separate the words in the sentence with Spaces according to the word segmentation model, and divide the sentence into one word (or word) unit; Dictionary filtering is used to remove English boundary characters and remove non-dictionary words from the corpus. After a series of processing of the corpus, we can get a relatively clean corpus that can be used for training, and then conduct model training.

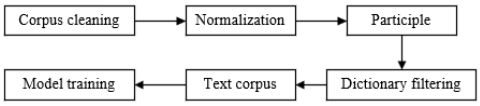

Fig. 2. The processing flow of Chinese corpus

3. Recurrent Neural Network Language Model

A typical recurrent neural network consists of three layers: input layer, hidden layer and output layer. The standard recurrent neural network (RNN) used in this paper, also known as Elman network, is easy to implement and train. After training on this RNN architecture, the probability of the current word w_i is expressed as:

$$p(w_i|h_i) = p_{rnn}(w_i|w_{i-1}, h_{i-1}) = p_{rnn}(w_i|h_i). \tag{4}$$

Where h_i represents all the context information of the current word in the statement, which is stored in the network storage layer in the form of a vector as a part of the network input when the next sample is trained.

It is assumed that at time t, the input word sample is $w(t)$, that is, the word vector of the current word, and the dimension is determined by the number of word samples $|V|$ in the corpus. The state $h(t)$ of the hidden layer is determined by the input current word vector $w(t)$ and the state of the hidden layer at the previous time, that is, the historical information $h(t-1)$. Through the connection between the hidden layer and the input layer, the state of the hidden layer at time $t-1$ is taken as a part of the input at time t. The output layer $y(t)$ represents the probability distribution information of the following words in the current history, and the number of nodes in the output layer is the same as the number of nodes in the input layer. The computational relationship between each layer is represented by the following formula:

$$Hidden \quad layer \quad input : x(t) = w(t) + h(t-1). \tag{5}$$

$$Hidden \quad layer \quad state : h_j(t) = f(\sum_i x_i(t)u_{ji}). \tag{6}$$

$$Output \quad layer \quad state : y_k(t) = g(\sum_j h_j(t)v_{kj}). \tag{7}$$

$$Sigmoid \quad activation \quad function : f(z) = \frac{1}{1+e^{-z}}. \tag{8}$$

$$Softmax \quad function : g(z_m) = \frac{e^{-z_m}}{\sum_k e^{-z_k}}. \tag{9}$$

softmax ensures that the probability distribution of the following words under the current word is reasonable, that is, $y_m(t) > 0$ for any word m, and $\sum_k y_k(t) = 1$. In the initialization setting of model parameters, the initial state $h(0)$ of the hidden layer is generally set to zero, or randomly initialized to a small value. The input word vector $w(t)$ is represented by a one-hot-vector. The number of hidden layer nodes is usually 100 to 1000, which is adjusted according to the size of specific training data. U, W, and V are weight matrices between layers, randomly initialized to smaller values. In the process of model training, the standard back propagation (BP) algorithm combined with stochastic gradient descend (SGD) is used to learn and update.

$$e(t) = desired(t) - y(t). \tag{10}$$

$$\theta(t) = \theta(t-1) - \alpha \nabla_{SGD}. \tag{11}$$

Where $e(t)$ represents the error vector at the output when each sample is processed, $desired(t)$ represents the probability value of the expected output in a particular context, and $y(t)$ is the actual output of the network. $\theta = u, w, v$ represents the corresponding gradient descent value, and the initial learning rate α of the network is 0.1, which is iteratively learned until the model converges.

4. RNN Model Adaptive

Language model plays an important role in improving the performance of speech recognition system. The adaptive language model is to compensate the information contained in the adaptive model and the missing information in the original model to get a more accurate model, so that the adaptive language model and the application environment can be matched to the greatest extent. In the case of a small amount of telephone tagging corpus, it is a good choice to use language model adaptive technology to reduce the language difference between the model and the recognition task, so as to adapt to the characteristics of different application environments and provide a more accurate language model for decoding speech recognition.

The structure of the online adaptive speech recognition system based on the RNN language model is shown in Figure 3, and the language model adaptive processing module is added to the end of the original system. When the RNN model is used to identify the system, the RNN language model is used to re-score the n-best list of the decoded word confusion network, which can get the one-best recognition result of the system under the RNN model. Although the recognition result of the system may not be completely correct, it can reflect the theme and language distribution of the recognition task to a certain extent. Therefore, the one-pass recognition results can be used as the model adaptive corpus, and after a series of processing such as word segmentation, the original RNN language model can be further trained and the parameters of the RNN model can be further updated. At this time, the model can learn new knowledge related to the recognition task, constantly adjust the probability of various language phenomena in the language model, better predict the real language distribution of the recognition task, and achieve model self-adaptation. The adaptive RNN model is re-scored on the speech n-best list to get a new language model score for each list. Then, the total score of each list is calculated according to equation (12), combined with the acoustic model score and penalty score information.

$$\lg L(s) = n \cdot wp + \sum_{i=1}^{n} asc_i + lms \sum_{i=1}^{n} \lg p_{mn}(w_i|h_i). \tag{12}$$

Where n is the number of words in the sentence. wp is the penalty part of the word. asc_i scores for word w_i acoustic model. lms is the model scale. $p_{mn}(w_i|h_i)$ represents the RNN language model score for each word. After the score of each list is calculated, the one with the highest score is selected as the optimal solution of the n-best list. The recognition rate of the system is calculated by comparing the newly obtained one-best list with the identification task annotation data.

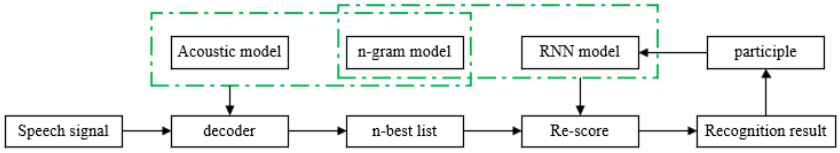

Fig. 3. Architecture of online adaptive speech recognition system based on RNN language model

5. Experiment and Analysis

In the experiment, Word Error Rate (WER) is selected as the evaluation standard. The lower the system word error rate, the better the performance of the language model is considered. The training data comes from the actual Chinese telephone channel voice label data provided by iflytek Voice Co., LTD., with a total of 16.5M, including 550k sentence text, including 4342k words, the verification set is 282k, and the test set is a list of 343300 sentences (100-best) decoding results of telephone speech with a size of 87k.

When RNN is used to train the model, the number of Hidden nodes and class of corpus are set to 100H-500C, 200H-100C, 500H-400C, 500H-500C, 600H-500C respectively, and the training of each model is completed. Then, the trained RNN model is used to re-score the language model of the 343300 sentences list, and the 3433 sentences with the highest score are selected as the result of one pass recognition. The RNN adaptive method above is used to retrain each RNN model. The system recognition rate of the adaptive model is tested. In this experiment, 3-gram model and Kneser-Neyback-off smoothing algorithm with good performance are used to train n-gram language model, which is constructed by SRILM toolbox.

In Table 1, the 3-gram model is taken as the baseline model, and the experimental results after RNN language model and RNN self-adaptation are respectively given. The experimental results after RNN self-adaptation and 3-gram model interpolation fusion are also given.

Table 1. Font sizes

Model	3-gram	RNN	RNN self-adaption	RNN self-adaption+3-gram
100H-500C	42.09	40.2	39.49	39.02
200H-100C	42.09	40.05	39.33	38.85
500H-300C	42.09	39.81	38.93	38.42
500H-400C	42.09	39.92	38.87	38.23
500H-500C	42.09	39.87	38.91	38.32
600H-500C	42.09	39.89	38.93	38.29

The experimental results are shown in Table 1. Compared with n-gram language model, using RNN modeling technology to train Chinese language model, WER of the system is effectively reduced, which indicates the superiority of RNN modeling technology and is suitable for Chinese language model training. After the adaptive training of the RNN model, the absolute WER of the system decreases by 1% compared with that of the RNN model system, and after the linear interpolation between the adaptive RNN model and 3-gram, the absolute WER of the system decreases by 1.5%, which shows the effectiveness of the adaptive algorithm proposed in this paper. After the RNN model is retrained, new knowledge related to the recognition task is learned, so that the model and the recognition task are matched to a certain extent. However, it can also be seen that due to the limited corpus, the system recognition rate of the original RNN model is not high, so there are some recognition errors in the first-pass recognition results of the system. After RNN self-adaptation, the model will also learn some wrong knowledge, which will affect the system recognition rate. In spite of this, the adaptive algorithm is still an effective algorithm in practical application because it is not large in time cost and can improve the performance of the model to a certain extent.

6. Conclusion

In order to solve the problem of differences between RNN generation model and recognition task, this paper proposes an RNN model adaptive algorithm, which does not integrate the traditional domain training language model and then interpolate with the general model. Instead, the algorithm takes the preliminary recognition results of speech as training corpus, reprocesses them and continues to train the RNN model. Since the text of the initial recognition results is not large, the model weight parameters can be updated quickly, and the probability of various language phenomena in the language model can be adjusted constantly, so that the adaptive RNN model and the recognition task can be matched to the greatest extent. Experiments show that the online adaptive algorithm of RNN language model proposed in this paper can effectively reduce the recognition error rate of the system. In addition, it can be seen from the experimental results that the first-pass recognition results of the system have a great impact on the performance of the adaptive model. Therefore, how to adjust the RNN training algorithm to improve the first-pass recognition rate of the system is the next research focus. It should be noted that several empirical parameters are used in the experiment. If the training conditions of the model change, these parameters should be adjusted accordingly.

7. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. This work was supported by the ”Research on language adaptive model based on Deep learning in future-oriented teaching scenarios”, project number: 25B413010.

References

1. Li J. Recent advances in end-to-end automatic speech recognition[J]. APSIPA Transactions on Signal and Information Processing, 2022, 11(1).
2. Li H, Teng L, Yin S. A new bidirectional research chord method based on bacterial foraging algorithm[J]. Journal of Computers, 2018, 29(3): 210-219.
3. Alharbi S, Alrazgan M, Alrashed A, et al. Automatic speech recognition: Systematic literature review[J]. IEEE Access, 2021, 9: 131858-131876.
4. Dhanjal A S, Singh W. A comprehensive survey on automatic speech recognition using neural networks[J]. Multimedia Tools and Applications, 2024, 83(8): 23367-23412.
5. Lin T, Li H, Yin S. Modified pyramid dual tree direction filter-based image de-noising via curvature scale and non-local mean multi-grade remnant multi-grade remnant filter[J]. International Journal of Communication Systems, 2018, 31(16).
6. Prabhavalkar R, Hori T, Sainath T N, et al. End-to-end speech recognition: A survey[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 32: 325-351.
7. Yu J, Li H, Yin S. New intelligent interface study based on K-means gaze tracking[J]. International Journal of Computational Science and Engineering, 2019, 18(1): 12-20.
8. Kim K, Wu F, Peng Y, et al. E-branchformer: Branchformer with enhanced merging for speech recognition[C]//2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023: 84-91.
9. Yao Z, Guo L, Yang X, et al. Zipformer: A faster and better encoder for automatic speech recognition[J]. arxiv preprint arxiv:2310.11230, 2023.
10. Sun Y, Yin S, Li H, et al. GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster[J]. IEEE Access, 2019, 7: 99254-99262.
11. Kim S, Gholami A, Shaw A, et al. Squeezeformer: An efficient transformer for automatic speech recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 9361-9373.
12. Jeon J, Lee S, Choe H. Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning[J]. Computers & Education, 2023, 206: 104898.
13. Fathullah Y, Wu C, Lakomkin E, et al. Prompting large language models with speech recognition abilities[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 13351-13355.
14. Weng Z, Qin Z, Tao X, et al. Deep learning enabled semantic communications with speech recognition and synthesis[J]. IEEE Transactions on Wireless Communications, 2023, 22(9): 6227-6240.
15. Yu J, Li H, Yin S L, et al. Dynamic gesture recognition based on deep learning in human-to-computer interfaces[J]. Journal of Applied Science and Engineering, 2020, 23(1): 31-38.
16. Hema C, Marquez F P G. Emotional speech recognition using cnn and deep learning techniques[J]. Applied Acoustics, 2023, 211: 109492.
17. Feng S, Kudina O, Halpern B M, et al. Quantifying bias in automatic speech recognition[J]. arxiv preprint arxiv:2103.15122, 2021.
18. Yu J, Lu Z, Yin S, et al. News recommendation model based on encoder graph neural network and bat optimization in online social multimedia art education[J]. Computer Science and Information Systems, 2024, 21(3): 989-1012.
19. Yin S, Li H, Laghari A A, et al. An Anomaly Detection Model Based on Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet of Everything[J]. IEEE Internet of Things Journal, 2024, 11(18): 29402-29411.
20. Ma P, Petridis S, Pantic M. Visual speech recognition for multiple languages in the wild[J]. Nature Machine Intelligence, 2022, 4(11): 930-939.

Biography

Jiangjiang Li is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.
Jiaxiang Wang is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.
Lijuan Feng is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.
Yachao Zhang is with the School of Electronics and Electrical Engineering, Zhengzhou University of Science and Technology. Research direction is computer application and AI.