Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230888

The application of machine learning algorithms in speech recognition error detection

Jiayi Yang

Computer Science and Technology, Northeast Forestry University, Harbin, 150000, China

2020222850@nefu.edu.cn

Abstract. The field of speech recognition technology has experienced a significant progress in recent years, with error detection being a crucial research domain. The present study offers a comprehensive overview of this area. Prior to the emergence of neural networks, Hidden Markov models (HMM) have been widely employed as a primary framework for speech recognition. However, this model only achieves local optimality, leading to its gradual replacement by neural network models, which have attracted considerable attention from researchers aiming to enhance their recognition performance. Various models, such as variant RNNs, bidirectional recurrent neural networks, and Fastcorrect2 have been developed. This paper introduces HMM, followed by a presentation of several neural network models, which entail a detailed description of their respective framework, principle, and idea. The variant RNN model is designed to enhance the recursive connection between input and output layers, while the deep bidirectional recurrent neural network model simulates the nonlinear relationship between input feature vectors and output labels using two models, namely bidirectional and deep. Additionally, the FastCorrect2 model enhances the voting effect of candidate words and the alignment algorithm. Finally, the study highlights the application of speech recognition error detection in everyday life, emphasizing the importance of speech recognition.

Keywords: error detection, speech recognition, NN, RNN, HMMs.

1. Introduction

Language is of great importance in the history of human development, as a tool for interpersonal communication and the exchange of ideas. Consequently, individuals continuously acquire languages throughout their lives, learning not only their native language but also additional languages to meet diverse needs by 2020, it's anticipated that more than 1.5 billion individuals will speak and acquire English as a second language globally [1]. Language learners, whether children or adults, will encounter inevitable obstacles, such as mispronunciation, grammatical errors, and semantic errors when acquiring a new language. Traditional methods for resolving these difficulties include soliciting assistance from family, friends, teachers, or seeking paid help through language testing organizations. However, these methods are often costly, inefficient, and time-consuming, potentially undermining the confidence of learners when the learning outcomes do not meet expectations. When the learning

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/16/20230888

results are not satisfactory, it can reduce people's confidence in the learning process, so it is especially important to find more effective alternatives. That's why more and more language learners are now looking for new ways to achieve better learning results. Artificial intelligence (AI) has played an important role in education in recent years, it can be considered as a time-saving and efficient way to learn a foreign language, and recently many investors are focusing on this field. And more and more language learning websites and applications are now using AI-related technologies, thereby AI is a new and efficient way to learn languages that are suitable for widespread use by language learners.

With speech recognition, computer vision, natural language processing, and other technologies becoming mature, artificial intelligence technology has advanced quickly in recent years. Numerous industry titans have increased their investments in AI research and development while also making their AI frameworks open source.so AI has emerged with some very important development history and development on nodes. In 1950, Alan Turing, the father of artificial intelligence, proposed the groundbreaking "Turing Test" to determine whether a computer meets the criteria for artificial intelligence; In 2011, IBM's Watson Deep Quiz System won the famous encyclopedic knowledge quiz TV show in the U.S., beating several outstanding human contestants and reaching the "Turing Test" closer; in May 2017, Alpha Go played against Ke Jie, the world's No. 1 Go champion, and won by a total score of 3-0, Alpha Go has reached the top human level. In 2020, OpenAI released GPT-3, the world's largest pre-trained language model, which is already approaching the human level in many practical language processing levels. Artificial intelligence also has some relevant research in language error correction, For example, Dong et al. introduced a hidden Markov model with a context-based deep neural network dubbed (CD)-DNN-HMM in 2011 that significantly outperformed the HMM-GMM system;[2] Errattahia et al. assessed and proposed a novel method for detecting errors that incorporates label dependency using a specialized RNN to further improve the error detection results in Automatic Speech Recognition (ASR); Gekhman et al. proposed to embed ASR word-level confidence scores into Transformer-based ASR error detectors using RED-ACE, which can lead to significant performance gains. The use of feature-based methods has been prevalent in the detection of errors in speech recognition. However, employing artificial intelligence techniques for this purpose represents a challenging and emerging field. Consequently, this paper aims to provide a comprehensive review of this domain.

This article is divided into several parts. The following content of this paper is organized as follows, the second part describes and analyzes the error detection methods in mainstream ASR; the third part will introduce the application scenarios of these error detection methods; the last part summarizes the whole study.

2. Method

2.1. Overview of the proposed method

This section will introduce various techniques applied to speech recognition error detection, The Hidden Markov Model (HMM) has traditionally been a popular approach for this task, but it has been increasingly supplanted by neural networks. Presently, researchers are placing greater emphasis on neural network approaches for speech recognition error detection.

In general, the research process of each method in this field includes several processes: design of relevant hardware, data and setup, model training, application of the model to a test set of results, comparison with model library, and draw conclusion as shown in Figure 1.

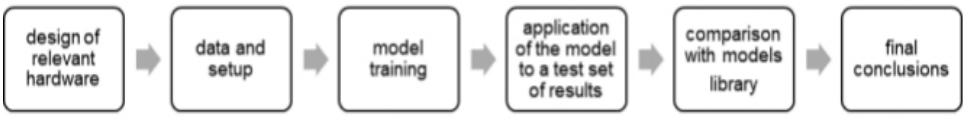

Figure 1. The procedure of speech recognition error detection.

DOI: 10.54254/2755-2721/16/20230888

2.2. Model

2.2.1. Hidden markov models (HMMs)

The early approach for speech recognition error detection was provided by the Hidden Markov Model (HMM), which offers a simple yet effective framework. The process of creating unobservable random sequences of states from a hidden Markov chain and then creating observed random sequences from individual states is described by hidden Markov models, which are probabilistic models concerning temporal sequences [3]. In general, HMM-based models can be divided into three modules: acoustic, articulation, and language models. Different modules have different roles, and each module is optimized independently with its optimization objective function. However, HMM also has the resulting drawback that such modularity achieves only local optimality. Thus, HMM models are gradually replaced by neural networks.

2.2.2. NN model

Markov models and neural networks are both statistical models that are graphically depicted, with the difference that neural networks are organized in parallel.Several types of neural networks can be used for speech classification, such as multilayer feedforward networks and Radial Basis Function Network trained using the backpropagation algorithm. In this model, data preprocessing is first performed on the relevant input items to ensure that the input to the neural network is a feature of each word, then the features are extracted, and the measure is estimated by a model library Then, the features are extracted, the measurement is estimated by metrics, and finally the recognition decision is made. With the research on neural networks in recent years, more advanced models have been utilized for error detection, such as convolutional network models, and recurrent neural networks.

2.2.3. Variational model

A variant of RNN, Variational RNN (V-RNN), is an evolution of Jordan RNN. Hannani et al. applied and evaluated this variant of RNN in ASR error detection [4]. ASR errors usually do not occur alone because unrecognized words end up generating the wrong sequence, and better classification results are achieved using V-RNN. Unlike traditional RNN models that are recursively connected only in the hidden layer, the V-RNN model performs recursive connections between the input and output layers. Errattahi et al. evaluated the performance of V-RNN by classification accuracy, precision, recall, and other metrics, and the final results showed that due to the label-dependent learning strategy in V-RNN, V-RNN achieved a significant increase in classification accuracy.

2.2.4. Deep bidirectional neural network model

Deep bidirectional recurrent neural networks (DBRNNs) model was proposed by Ogawa et al. for the detection of uncommon labels, which has a good conditional random field (CRF) and has high error detection performance [5]. This model is an RNN model with deep and bidirectional structure, where the deep bidirectional model (BRNN) can be propagated through the forward and backward activated vectors, in the hidden layer, to complete the context of the feature vectors considering the input throughout the time step and thus the output labels at the current time. The hidden layers are stacked in the forward and backward directions, respectively, in the deep structure, i.e., DBRNN, in order to model the nonlinear relationship between the input feature vectors and the output labels and to show that the deep structure has a significant impact.

2.2.5. Deep bidirectional neural network model

The FastCorrect2 model was proposed by Feng et al. This model exploits the voting effect of multiple candidate words through an alignment algorithm to help better detect and correct mislabeling. This model uses multiple candidates generated by ASR beam search to help the error correction, and it makes several modifications to the previous non-autoregressive error correction model to make it suitable for inputs with multiple candidates [6]. At the same time, FastCorrect 2 model also surpasses

DOI: 10.54254/2755-2721/16/20230888

the single-candidate non-autoregressive correction model in terms of correction accuracy, proving the value of multi-candidate error correction. The model is structured in four parts, encoder, duration predictor, candidate predictor, and decoder, some of them have various innovations. For example, the addition of a Pre-Net to the Transformer encoder takes full advantage of the voting effect; the combination of the encoder output and the original encoder input (different for each candidate) on each location ensure that it is more discriminative; The training model set is pre-trained to enable the model to select the simplest candidate for calibration.

3. Application and discussion

3.1. Application in computer-assisted language learning

ASR is used in various fields, the most well-known of which is its application in the field of computer-assisted language learning. ASR technology is often used to improve the pronunciation of foreign language learners by recording their speech, performing acoustic analysis, and comparing it with native language samples. However, this application also encounters several hurdles, such as disfluency, grammatical errors, and mispronunciation in the speech of foreign language learners. Numerous end-to-end systems have been developed to address disfluency detection and attain high performance [7]. However, further research is still needed for the prediction of language and the improvement of error recognition scores.

3.2. Application in smart classroom

The utilization of ASR technology in smart classrooms has also received much attention, owing to advancements in big data, IoT technology, and deep learning frameworks that enhance cloud computing. This enables human-computer interaction and facilitates intelligent systems in comprehending user commands [8]. At the same time the openness of the platform can increase the training scenarios in the classroom and make up for the lack of applications in classroom teaching. However, the current development of cloud computing technology is not mature enough, resulting in the inability to make ASR better applied in the cloud, as well as the difficulty of uploading terminal data to the cloud for storage. The future will be dedicated to dealing with the relationship between capital investment and hardware integration while concurrently increasing the adoption of cloud computing.

3.3. Application in medical field

Nowadays, due to the development of artificial neural network models, the application of ASR has also started to enter the industrialization and ecological stage [9]. In recent years ASR has also started to be applied in the medical field [10], where doctors can enter cases by voice input to improve the efficiency of entry; or enter electronic cases so that when patients want to make secondary inquiries, they do not need to be reviewed by doctors. However, there are fewer studies on disease assisted diagnosis based on ASR, and it can only be applied to a few diseases, and there are huge challenges in accuracy rate. Future research is also continuing, taking into account the actual situation as well as the needs.

4. Conclusion

In summary, this paper provides an overview of error detection in speech recognition in recent years, an important module in speech recognition where error detection is concerned. This paper summarizes the five models mainly applied in this field and the functions of each model. Among them, Hidden Markov Models are the early techniques that are heavily used in error detection systems. Later, with the technological development of neural networks, various evaluation metrics about error detection have been substantially improved and error detection systems have been optimized to a great extent. These models are now heavily used in manufacturing, language learning, and even medicine, as well as in all aspects of people's lives. In the future, there will be more improved models based on neural

DOI: 10.54254/2755-2721/16/20230888

networks to further improve the speed and accuracy of error detection in speech recognition, so that speech recognition can play a better role in people's lives and make their lives more convenient in all aspects.

References

[1] Council B 2013 The English Effect Aug Research Report.
[2] Dahl G E et al 2011 Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition IEEE Trans Audio Speech Lang Process 20 pp 30–42
[3] Mor B et al 2021 A Systematic Review of Hidden Markov Models and Their Applications. Arch Computat Methods Eng 28 1429–1448 https://doi.org/10.1007/s11831-020-09422-4
[4] Hannani A E 2019 Incorporating label dependency for ASR error detection via rnn Procedia Computer Science Retrieved April 18 2023 from https://www.academia.edu/93987355/Incorporating_label_dependency_for_ASR_error_detection_via_RNN
[5] Ogawa A and Hori T 2015 ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4370-4374 IEEE
[6] Leng Y Tan X Wang R et al. 2021 Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition arXiv preprint arXiv:2109.14420
[7] Mateiu T 2022 End-to-End Disfluency Detection in Automatic Speech Recognition for Second Language Learners Aalto University, Master’s Programme in Computer, Communication and Information Sciences
[8] Lijun H and Jing J 2022 Smart Speech Recognition System for Chinese Language Learning Enhancement Scientific Programming vol. 2022 Article ID 1701474 11 pages https://doi.org/10.1155/2022/1701474
[9] Zhao M et al 2022 Research on the application of intelligent speech recognition technology in medical field Chinese Modern Physicians (28) pp 108-112
[10] Vajpai J and Bora A 2016 Industrial applications of automatic speech recognition systems International Journal of Engineering Research and Applications 6(3) pp 88-95