An Acoustic Recognition End-to-end System on Under-resources Languages

Chun-Hao Wang
Department of Computer Science and Information Engineering
National Taiwan University
chuchuhao831@gmail.com

***Abstract*—With the rapid development of sequence learning techniques, the state-of-the-art speech recognition system have a great performance on English speech corpus with almost human ability. Compare to others under-resourced language which is less language-specific knowledge or less hand-labeling data is still an active field for research and still have no good performance. Instead to find a system can make a good performance on some language recognize problem, we present a end-to-end system that directly translate the acoustic sequence to label sequence without any language-specific knowledge so that we can adapt on any language without much resources. An experiment on the Cebuano speech corpus demonstrates its ability indeed learning some pattern from data.**

***Index Terns*—** under-resources language, minority languages, connectionist temporal classification (CTC), sequence learning, speech and language resources, automatic speech recognition (ASR), cross-lingual acoustic modeling.

I. Introduction

Speech recognition problem takes advantage of recent advance in algorithms and computer hardware, using the artificial neural networks (ANNs) techniques acquire a good performance on the task. According to the Grave's research[2],*a end-to-end system achieves a word error rate of 27.3% on the Wall Street Journal corpus with no prior linguistic information, 21.9% with only a lexicon of allowed words, and 8.2% with a trigram language model. Combining the network with a baseline system further reduces the error rate to 6.7%*. Not only to this end-to-end system acquire that perfect result, there also some graphical models such as hidden Markov Models (HMMs; Rabiner), conditional random field (CRFs; Lafferty) hybrid with ANNs and the aggregation of different models with additional language-specific feature, all of them do the great job. While all of these approaches performed excellent, but they usually have to use lots of language resources which from large hand-labeling acoustic sequence to label sequence pair to some knowledge like phoneme and feature theory except the Grave's RNN-CTC system. Grave's system can achieves low error rate in English corpus without prior linguistic information, and we assume such characteristic which regardless of linguistic information let us be able to apply the system on some other language.

So far, the speech recognition problem on under-resourced language is still an active field for research, and we expect that if we could apply a system without the language-specific knowledge (like the Grammer, Grapheme and Phonology) and less labeling data then we create a system as the speech-to-text machine on all the language in the world. We believe we can break the language barrier for human communication.

To define clarity, we seem this as a sequence learning problem which labeling unsegmented sequence data or to say unsupervised learning the temporal classification[4]. Input a time various speech sequence data and output a sequence of label, for example, we take the filter-bank extract acoustic features from the speech signal with a fixed frame size as input and turn them into a sequence of labels which is English Character as output.

In our experiment we use bidirectional long short-term memory as our neural unit with their powerful, general mechanism for modelling time series and connectionist temporal result (CTC) output to let the RNN-alone sequence learning become possible.

The next section introduce bidirectional long short-term memory (BLSTM) and connectionist temporal classification (CTC) and briefly describe how them be trained, output representation and build our recognition system. Section 3 describe the Cebuano speech corpus we use in our experiment. Section 4 present our experiment result and discuss them. Section 5 makes a conclusion and provide list of further work we are on the process to solve this problem.

II. Recognition system architecture

Artificial neural network is a box that take an input and return an output target. The architecture inside is a single or multiple layer with each layer consist a list of neurons. All these neurons and the weighted connection between the layer build that box. To let our box can learn the ability to response the correct output a we want, we define a proper objective function that represent the score of it's learning outcome. We try to adjust the connection weights to get the better score. Here, we are not going the detail of mathematical formalism for the reason that all of them can be found in the reference[4], what here is just some my own reviews of this box architecture. Last we combine the CTC output layer to the network and train them with the log-likelihood to let the system learn from sequences to sequences not only frames to frame. The reason we use bidirectional long short-term memory as our neurons is explain and following subsection and what CTC do in section 2-B.

A. Bidirectional long short-term memory

A standard recurrent neural network (RNN) computes the hidden vector sequences and output vector sequence by an input sequence, with it's cyclic architecture in neural connection it reveal the ability to memorize the temporal information for a better modeling. However researcher have found that Long Short-Term Memory (LSTM) architecture [5], which uses purpose-built memory cells to store information, is better at finding and exploiting long range context and solve some training problem with standard RNN[6]. Since the standards RNNs process sequence in temporal order, they do not use the future context which is obviously important in sequence learning system. There are two method to solve this, one is to add a time-window of the future context to the network input, for example, at time t, in standard RNNs, we take only data at time t, but with time-window, we concatenate t and t+1 data as input, which double the dimension and use the future data. However, as well as unnecessarily increasing the dimension of input weights, this suffer from the intolerance of distortions. Bidirectional recurrent networks (BRNN)[7] offer a more elegant solution. The basic idea of BRNNs is to present each training sequence to forward and backward to two separate recurrent hidden layers, both of which are connected to the same output layer. Briefly speaking, standard RNNs train temporally from time beginning of the sequence to the end, on the contrary, BRNNs train not only forward time ordering but also the backward time ordering to get the future context. Last we use the deep bidirectional LSTM[8] which means we take all advantage from the BRNNs, LSTM and deep structure in Artificial Neural Network as our system to do the speech recognition task.

B. Connectionist temporal classification

Although the BLSTM acquire a huge success on frame-level classifier in speech recognition, that still not solve problem we want, directly ???

III. Cebuano data processing

Cebuano is an Austronesian language spoken in the Philippines by about 20 million people. It has the largest native language-speaking population of the Philippines despite not being taught formally in schools and universities. There are some linguistic research about the language but we dose not use in this project. The reason we choose Cebuano is that we know nothing about the Cebuano expect a 3 hours speech data with their target label. Each data pair represent a speech slice consist a sequence of speech data and a correspondent sequence of labels(character). The following section 3-A we describe how we process the speech data, ans some observation in Cebuano target and what we handle them in section 3-B.

A. Processing speech data

There's two ways we process the speech to our system input, one is 30-dimension mel-frequency cepstral coefficients(MFCCs) and 90-dimension DNN trained bottleneck feature. Both of these two front-end feature extract we use in our experiment. The MFCCs approach is the traditional ways to do the front-end job. The bottle-neck feature is a MLP extract feature[9][10].

B. Observation in Cebuano target

We discuss the observation from the training and testing target file, which are sequences of character as the training target, in two aspect: characters and words. On the character view, there are 32 different characters includes English alphabet (a-z) and specific character (” ”, ””””, ”-”, ”_”, ”.”, ”?”), the ” ” represent the space, ”.” is the blank (for CTC output) and ”?” means some noise that have no meaning. We don't know the usage of all other specific character, but we think they are important and should be conclude. These characters we names them the labels. On the words side, we do not know exactly what is a word in Cebuano but simply extract a word when a sequence of label meet the *space*. There are 3XXX different words in training data and 3XXX different words in testing data. But, the overlap of them only got 1XXX words.

IV. Experiment results

Architecture we use in our experiment have two different input model, one is The prediction network All networks were trained with RMSprop using s learning rate $10^{-}4$ and a momentum of 0.9. Gaussian weight noise[**] with a standard deviation of 0.075 was injected during training to reduce overfitting.

The label error rate is not good enough to do more works

Experiment result get here

V. Conclusion

The conclusion goes here.

Acknowledgment

Special thanks to our professor Hung-Yi Lee, who open a course named *Machine Learning and having it deep and structured (2015,Spring)* in National Taiwan University, all of my knowledge about neural network are learning from the class and project also inspired by the final project in that course which is build an ASR system on English. Another important role is Chan-Hou Chen, who enroll the course with me and propose to do an end-to-end system on our final project and implement the algorithm in python. Thanks to Michael Shell who create this Latex template.

References

[1] L. Besacier, *Automatic Speech Recognition for Under-Resourced Languages: A Survey*

[2] A. Grave, *Towards End-to-End Speech Recognition with Recurrent Neural Networks*

[3] A. Grave, *Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Network*

[4] A. Grace, *Supervised Sequence Labelling with Recurrent Neural Network*

[5] S. Hochreiter, *Long Short-term memory*

[6] R. Pascanu , *On the difficulty of training recurrent neural networks*

[7] M. Schuster , *Bidirectional Recurrent Neural Networks*

[8] A. Grace, *Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition*

[9] F. Grezl, *Probabilistic and Bottle-Neck Feature for LVCSR of Meetings*

[10] S. Tomas, *Deep Neural Network Features and Semi-supervised Training For Low Resource Speech Recognition*

[11] Jim, *An Analysis of Noise in Recurrent Neural Networks Convergence and Generalization*