ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414935

A UNIVERSAL BERT-BASED FRONT-END MODEL FOR MANDARIN TEXT-TO-SPEECH SYNTHESIS

Zilong Bai, Beibei Hu

Ajmide Media, P.R. China

{baizilong, hubeibei}@ajmide.com

ABSTRACT

The front-end text processing module is considered as an essential part that influences the intelligibility and naturalness of a Mandarin text-to-speech system significantly. For commercial text-to-speech systems, the Mandarin front-end should meet the requirements of high accuracy and low time latency while also ensuring maintainability. In this paper, we propose a universal BERT-based model that can be used for various tasks in the Mandarin front-end without changing its architecture. The feature extractor and classifiers in the model are shared for several sub-tasks, which improves the expandability and maintainability. We trained and evaluated the model with polyphone disambiguation, text normalization, and prosodic boundary prediction for single task modules and multi-task learning. Results show that, the model maintains high performance for single task modules and shows higher accuracy and lower time latency for multi-task modules, indicating that the proposed universal front-end model is promising as a maintainable Mandarin front-end for commercial applications.

Index Terms— text-to-speech, multi-task learning, front-end, BERT

1. INTRODUCTION

The front-end text processing system plays an important role that influences the intelligibility and naturalness in a Mandarin text-to-speech (TTS) system [1, 2]. The typical Mandarin front-end is usually designed as a pipeline-based structure that consists of a series of individual components, such as, polyphone disambiguation (PD), text normalization (TN), prosodic boundary prediction (PBP), Chinese word segmentation (CWS), part-of-speech (POS), etc. Each component in the front-end can be addressed by a rule-based algorithm or a data-driven method. Application of multi-task learning (MTL) and fine-tuning a pre-trained model showed a mount of impressive results in front-end text processing tasks [3, 4]. Nevertheless, a pipeline-based front-end with complex structure also brings several problems including error propagation, inference latency, and misalignment in optimization [1]. Moreover, each component is modeled separately, which increases the complexity of the system and reduces maintainability. To address the above issues, sequence-to-sequence (seq2seq) models for unified front-ends were suggested and achieved promising results [5, 6]. However, the sequence output directly from a seq2seq model can lead to unrecoverable errors [5]. Mandarin corpuses with multi labels or tags are more difficult to obtain and maintenance than corpuses with single label.

In this paper, we propose a universal BERT-based model that can be used for various tasks in the Mandarin front-end without changing its architecture. The input of the proposed model has two branches, one for classification tasks and the other for sequence tagging tasks, where PD, TN, PBP, CWS, and POS are included. The two branches can be shared for a group of associated sub-tasks with identified masks and optimized alternately for MTL. We fine-tuned the proposed BERT-based model with PD, TN, and PBP separately and with their combinations for MTL. The evaluation results show that, the proposed model maintains high performance for single task modules. The module with multi-task model for PD and TN boosts the accuracy by 0.35% and reduces time latency by 25.6% compared to those of our pipeline-based module, indicating the prospect of multi-task modules for Mandaring front-ends. In conclusion, the proposed model is promising as a maintainable Mandarin front-end for commercial applications.

2. BACKGROUND

In this section, we briefly review the three tasks used to train the proposed front-end model, including PD, TN, and PBP. All of these tasks benefiting from BERT pre-training inspired us to adopt a BERT-based feature extractor for the model.

2.1. Polyphone disambiguation

For Mandarin, the target of a grapheme-to-phoneme (G2P) system is to transform characters to their corresponding phoneme representation. However, one character could have a fixed or several candidates of pronunciations in terms of different context, which brings more challenges. The goal of a PD system is to addresses this homograph problem. Researches on PD include rule-based algorithms [7-9] and data-driven methods[10-12]. Rule-based methods require linguistic experts to produce an

elaborate text-analysis system and a robust dictionary, which is complex and expensive. In contrast, data-driven methods can extract statistical knowledge from annotated data in a relatively simpler way and outperform the rule-based approaches. Recently, pre-training and fine-tuning methods, such as BERT [13], GPT [14] and MASS [15], were leveraged in PD and achieved state-of-the-art accuracy, implying their promising application in PD.

2.2. Text normalization

TN is a process that transform non-standard words (NSWs) to spoken-form words (SFWs) for disambiguation. In chinese, for example, number expressions, time expressions and special symbols are major NSWs that need to be transformed to Chinese characters. The earliest Mandarin TN systems used manual rules based on keywords and regular expressions, which led to difficulties to improve the performance on general cases. For data-driven methods, maximum entropy (ME) [16, 17] and neural networks[6, 18] were employed for normalizing hard cases in TN. However, since these models were trained on limited annotation data, it is difficult to achieve satisfactory generalization performance. Recently, Zhang et al [19]. proposed a bybrid TN system that combined a BERT-based model with a rule based model. This model benefited from BERT pre-training and improved the accuracy on NSW-SFW transformation in sentence-level obviously.

2.3. Prosodic boundary prediction

Prosody structure plays an important role in Mandarin speech synthesis, in terms of both naturalness and intelligibility. In contrast to English, Mandarin is a kind of continuous writing language, so the prosody structure of Mandarin is more complex than that of English. Typical prosody prediction methods include rule-based models and statistical models like CRF [20], and RNN [21]. Recently, Talman et al. demonstrated that BERT-based model had implicitly learned syntactic or semantic information relevant for the PBP and outperformed the other models [22].

3. THE UNIVERSAL BERT-BASED FRONT-END MODEL

In this work, PD and TN were considered as classification problems and PBP was regarded as a sequence tagging problem. As shown in Fig. 1, the proposed model is constructed with a common BERT-based feature extractor following two separate linear neural networks as classifiers. The classifier of Task1 is shared by all sub-tasks for classification and the classifier of Task2 is shared by all sub-tasks for sequence tagging. Because each classifier can be shared for a group of associated sub-tasks with identified masks, new tasks can be easily introduced into the model by modifying the dimensions of the masks and classifiers instead of changing the model architecture. More details of the model are described as follows.

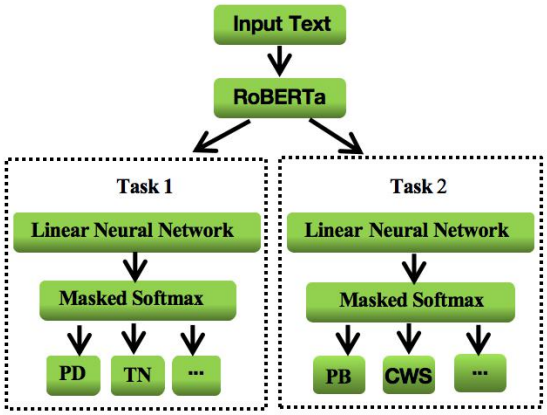

Fig. 1. Network architecture of the proposed model.

3.1. Contextual representation

BERT is a transformer model pre-trained on a large amount of unlabeled data and generates enriched contextual representation, showing its promising application in TTS front-end. Therefore, we introduced a pre-trained RoBERTa for generating contextual representation of input sentences and fine-tuned it to boost the performance. Each character in the input sequences was embedded as a fixed-length feature vector to accumulate the forward and backward context information as the contextual representation for downstream tasks.

3.2. Cross entropy loss with a mask

For PD and TN sub-tasks, each polyphonic character or NSW has a fixed number of classes corresponding to the candidate pronunciations or patterns, which inspired us to leverage a mask in the softmax classification to reduce influences of different characters on each other. The cross entropy loss with mask can be described by the following formula [3]:

$$\mathcal{L}(\theta) = -\sum_{(x,y)\in\mathcal{D}}\sum_{w\in W^x}\sum_{k\in\mathcal{C}_w}\mathbf{1}\{y_w = k\} \times logP(y_w = k|x;\theta), \quad (1)$$

where θ and D are the model parameters and training data respectively, W^x is the set of indices of the polyphonic words or NSWs in a training sentence x, y_w is the classification outputs of the corresponding polyphonic words or NSWs, C_w is the set of classes for the candidate pronunciations or patterns, **1** is the indicator function.

Similarly, each classifiers was shared by a group of associated sub-tasks using identified masks which masked the irrelevant classes and only calculating the loss of the candidate classes. The loss with class mask is:

$$\mathcal{L}(\theta) = -\sum_{(x,y)\in\mathcal{D}}\sum_{w\in T^x}\sum_{k\in\mathcal{C}_c}\mathbf{1}\{y_w = k\}\times logP(y_w = k|x;\theta), \quad (2)$$

where T^x is the set of indices for the corresponding sub-task in a training sentence x, and *Cc* is the set of candidate classes for the corresponding sub-task.

3.3. Multi-task learning

MTL is a promising machine learning paradigm that aims to improve the generalization performance by leveraging training signal contained in related tasks [23]. This is typically done by training a single neural network for multiple tasks jointly or alternately. To better illustrate generality of the proposed model, we fine-tuned a multi-task model for MTL where Task1 and Task2 were optimized alternately.

The proposed model consists of a common BERT-based feature extractor following two linear classifier output layers and two masked softmax layers for classification and sequence tagging. Task1 is the classification task that joins its sub-tasks by mixing their training data uniformly and so do sub-tasks in Task2 for sequence tagging. We selected Task2 as the reference task with parameter updating frequency of 1. The parameter updating frequency coefficient β for Task2 was then set as:

$$\beta = \frac{N_1}{N_1+N_2}, \quad (3)$$

where N_1 and N_2 are the total number of samples for training Task1 and Task2 , respectively. This configuration ensured all samples in the training dataset be used once in each epoch.

3.4. Hybrid TN Module

To fully leverage the combined advantages of rule-based models and neural network models, we introduced a hybrid TN module. The NSWs were first extracted from the input text using regular expressions. Then the system performed a priority check on the NSWs and transformed the matched NSWs to SFWs by a rule-based modle. All of the remaining patterns were passed through a neural network model to be classified into the right classes that concern to corresponding post processing methods to accomplish the conversion.

4. EXPERIMENTS

4.1. Dataset

As shown in Table 1, four internal datasets were used in the experiments for the corresponding tasks. Each dataset was labeled by linguistic experts with rich experience and double-checked to ensure the consistency and accuracy.

SET_PD: mandarin polyphone disambiguation dataset, which contains 136 frequently used polyphonic characters with their 287 corresponding pronunciations. There is only 1 polyphonic character per sample for classification.

SET_TN: mandarin text normalization dataset. There are 5 patterns and 14 classes in SET_TN. The patterns are digit or symbol related, and patterns like abbreviations are not included. In each sample, only one designated pattern was labeled for classification.

SET_PBP: Prosodic Boundary Dataset. The PBP task was treated as a character sequence tagging task, so each character in a sentence from SET_PBP was classified into a class that represents the corresponding level of a prosodic boundary.

SET_GOLDEN: An internal golden test set. The dataset includes 11140 pairs of sentences that were used to evaluate the inference accuracy in sentence-level in terms of TN and PD. The sample format like the followings: '最终的比分是5:3 (5:3 was the final score)' and '最终的[de5]比分[fen1]是五比三'.

	Training set	Evaluation set
SET_PD	89296	22323
SET_TN	10898	2736
SET_PBP	9574	2393
SET_GOLDEN	0	11140

Table 1. Number of sentences for training and evaluating in four internal datasets used in corresponding sub-tasks.

4.2. Training and evaluation

We adopted RoBERTa-wwm-ext (https://github.com/chineseGLUE/chineseGLUE) as the pre-trained model to extract the contextual representation, which consists of 12 Transformer layers, with 768 hidden dimension, 12 attention heads, and about 108 millions of parameters in total. In the first two epochs, only the two classifiers were trained with a learning rate of 1e-3, keeping parameters in RoBERTa frozen. In the remaining epochs, all parameters in the model were updated with a fixed learning rate of 2e-5.

In this work, we fine-tuned the proposed BERT-based model with PD, TN, and PBP separately and with their combinations for MTL. Then we performed evaluation and compared their performance by F1 score. Time latency and inference accuracy in sentence-level of a multi-task model with PD and TN were calculated to demonstrate the advantages of muti-task modules compared to single task modules in a Mandarin front-end.

Our server is equipped with a CPU of Intel Xeon Gold 6152 and memory capacity of 254 GB. All experiments were carried out on an NVIDIA Tesla V100 GPU with video memory of 32GB.

4.3. Results

We first evaluated the performance of the neural network models using evaluation set of SET_PD, SET_TN, and SET_PBP which are demonstrated in Table 1. Table 2 shows F1 scores of models trained with PD, TN and PBP separately and with their different combinations for MTL. All of the single task models and multi-task models maintained high performance in F1 score, which implies the generality and expandability of the model. The neural network structure can be easily shared by various of tasks, indicating its excellent excellent maintainability. Compared to the single task models, all multi-task models outperformed in TN, which confirms both PD and PBP are potential effective auxiliary tasks for TN. Inter connections of knowledge between the tasks may be responsible for this result. As for PBP, F1 scores of PD_PBP and PD_TN_PBP were slightly lower than that of PBP_Only, while comparable F1 scores between PBP_TN and PBP_Only were observed. In our datasets, the training data of PD was considerably larger than that of PBP, which decreased the optimization frequency of PBP. This may account for the lower F1 score in PD_PBP and PD_TN_PB.

	F1 scores in Task1		F1 score in Task2
	PD	TN	PBP
PD_Only	0.9875	~	~
TN_Only	~	0.9862	~
PBP_Only	~	~	0.9144
PD_TN	0.9873	0.9873	~
PBP_TN	~	0.9883	0.9134
PD_PBP	0.9867	~	0.9069
PD_TN_PBP	0.9867	0.9882	0.9053

Table 2. F1 scores of models trained with PD, TN, and PBP separately and with their different combinations for MTL.

To further demonstrate the advantages of a multi-task module based on the proposed model, time latency and inference accuracy in sentence-level were checked by SET_GOLDEN.

Apart from single task modules and a multi-task module for PD and TN, we constructed a pipeline-based with PD_Only and TN_Only. Table 3 depicts time latency of different modules for PD and TN. We define time latency as time interval between input time and output time. In our system, time latency consists of data processing time and inference time in neural networks. Though the data processing time in the multi-task module is longer than that of single task modules, it still less than the data processing time of the pipeline-based module.

	Data Processing Time (s)	Neural Network Inference Time (s)	Total Time Latency (s)
PD_Only	25.6	21.7	47.3
TN_Only	11.1	12.3	23.4
Pipeline-based module	36.7	34.0	70.7
PD_TN	30.5	22.1	52.6

Table 3. Time latency of different modules for PD and TN.

For each sample with multiple inference tasks, the BERT-based feature extractor runs only once and is shared by all sub-tasks, which decreases the inference time significantly. Compared to the pipeline-based module, the multi-task module reduced the data proceding time, neural network inference time and total time latency by 16.9 %, 35.0 % and 25.6%, respectively. The main contribution on reduceing time latency came from neural network inference time.

	Inference Accuracy
Pipeline-based module	0.9649
PD_TN	0.9684

Table 4. Inference accuracy of the constructed pipeline-based module and the multi-task module on sentence-level in terms of PD and TN.

Table 4. shows inference accuracy of the sctructed pipeline-based module and the multi-task module on sentence-level in terms of PD and TN. Benefit from MTL, the multi-task module boosts the accuracy by 0.35% compared to the constructed pipeline-based model. This result indicates the prospect of multi-task modules for Mandaring front-ends.

5. CONCLUSIONS

In this paper, we propose a universal BERT-based model that can be used for various tasks in the Mandarin front-end without changing its auchitecture. We fine-tuned the proposed BERT-based model with PD, TN, and PBP separately and with their combinations for MTL. The evaluation results show that, the proposed model maintains high accuracy for separate tasks in Mandarin front-ends, which implies the excellent generality and expandability of the model. The multi-task modules benefit from MTL and show better performance than that of the pipeline-based modules in Manderin front-end, indicating the prospect of multi-task modules for Mandaring front-ends. In conclusion, the proposed model is promising as a maintainable Mandarin front-end for commercial applications.

6. REFERENCES

[1] J. Pan ,X. Yin, Z. Zhang, S.Liu ,Y. Zhang, Z.Ma, and Y. Wang, "A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis," in International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 6689-6693.

[2] H. Pan, X. Li, and Z. Huang, "A Mandarin Prosodic Boundary Prediction Model Based on Multi-Task Learning," in Interspeech, 2019, pp. 4485-4488.

[3] H. Sun, X. Tan, J. Gan, S. Zhao, D. Han, H. Liu, T. Qin, and T.Y. Liu, "Knowledge Distillation from Bert in Pre-Training and Fine-Tuning for Polyphone Disambiguation," in Automatic Speech Recognition and Understanding Workshop, IEEE, 2019, pp. 168-175

[4] D. Dai1, Z. Wu, S. Kang, X. Wu3, J. Jia, D. Su, D. Y, and H. Meng, "Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-trained BERT," in Interspeech, 2019, pp. 2090-2094.

[5] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural models of text normalization for speech applications," Computational Linguistics, vol. 45, no. 2, pp. 293-337, 2019.

[6] C. Mansfield, M. Sun, Y. Liu, A. Gandhe, and B. Hoffmeister, "Neural text normalization with subword units," in Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, vol. 2, pp. 190-196,.

[7] H. Zhang, J. Yu, W. Zhan, and S. Yu, "Disambiguation of Chinese Polyphonic Characters," in International Workshop on MultiMedia Annotation, 2001, vol. 1, pp. 30-1.

[8] Z. Zirong, C. Min, and C. Eric, "An Efficient Way to Learn Rules for Grapheme-to-Phoneme Conversion in Chinese," in International Symposium on Chinese Spoken Language Processing, 2002, pp. 59-62.

[9] F. L. Huang, "Disambiguating Effectively Chinese Polyphonic Ambiguity Based on Unify Approach," in International Conference on Machine Learning and Cybernetics, 2008, vol. 6, pp. 3242-3246.

[10] C. Shan, X. Lei, and K. Yao, "A Bi-directional LSTM Approach for Polyphone Disambiguation in Mandarin Chinese," in International Symposium on Chinese Spoken Language Processing, IEEE, 2017, pp. 1-5

[11] F. Z. Liu and Y. Zhou, "Polyphone Disambiguation Based on Maximum Entropy Model in Mandarin Grapheme-to-Phoneme Conversion," Key Engineering Materials, vol. 480-481, pp. 1043-1048, 2011.

[12] J. Liu, W. Qu, X. Tang, Y. Zhang, and Y. Sun, "Polyphonic Word Disambiguation with Machine Learning Approaches," in 2010 Fourth International Conference on Genetic and Evolutionary Computing , 2010, pp. 244-247.

[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirec- tional transformers for language understanding," arXiv preprint arXiv: 1810.04805, 2018.,

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/language understanding paper.pdf, 2018.

[15] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation," arXiv preprint arXiv: 1905.02450, 2019.

[16] Y. Jia, D. Huang, W. Liu, Y. Dong, S. Yu, and H. Wang, "Text normalization in mandarin text-to-speech system," in International Conference on Acoustics, IEEE, 2008, pp. 4693-4696.

[17] T. Zhou,Y. Dong,D. Huang, W. Liu, and H. Wang, "A three-stage text normalization strategy for mandarin text-to-speech systems," in International Symposium on Chinese Spoken Language Processing. IEEE, 2008, pp. 1-4.

[18] R. Sproat and N. Jaitly, "Rnn approaches to text normalization: A challenge," arXiv preprint arXiv: 1611.00068, 2016.

[19] J. Zhang, J. Pan, X. Yin, S. Liu, Y. Zhang, Y. Wang, and Z. Ma, "A hybrid text normalization system using multi-head self-attention for mandarin," in To Be Submitted to International Conference on Acoustics, 2020

[20] Y. Zheng, J. Tao, Z. Wen and Y. Li, "BLSTM-CRF Based End- to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in A Text-to-Speech Front-End," in Interspeech, 2018, pp. 47-51.

[21] Z. Ying and X. Shi, "An rnn-based algorithm to detect prosodic phrase for chinese tts," in 200 International Conference on Acoustics, IEEE, 2001, vol. 2, pp. 809-812.

[22] A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, M. Vainio, "Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations," arXiv preprint arXiv:1908.02262, 2019.

[23] H. M. Alonso and B. Plank, "When is multitask learning effective? Semantic sequence prediction under varying data conditions," arXiv preprint arXiv:1612.02251, 2016.