INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Mandarin Prosodic Boundary Prediction Model Based on Multi-Task Learning

Huashan Pan, Xiulin Li, Zhiqiang Huang

Databaker (Beijing) Technology Co., Ltd, Beijing, China
{panhuashan, lixiulin, huangzhiqiang}@data-baker.com

Abstract

In this paper, we propose a mandarin prosodic boundary prediction model based on Multi-Task Learning (MTL) architecture. The prosody structure of mandarin is a three-level hierarchical structure, which contains three basic units--Prosodic Word (PW), Prosodic Phrase (PPH) and Intonational Phrase (IPH) [1]. Previous studies usually decompose mandarin prosodic boundary prediction task into three independent tasks on these three unit boundaries [1-4]. In recent years, with the development of deep learning, MTL has achieved state-of-the-art performance on many tasks in Natural Language Processing (NLP) field [5-7]. Inspired by this, this paper implements an MTL framework with Bidirectional Long-Short Term Memory and Conditional Random Field (BLSTM-CRF) as the basic model, and takes three independent tasks of mandarin prosodic boundary prediction as sub-modules for PW, PPH and IPH individually. Under the MTL architecture, the three independent tasks are unified for overall optimization. The experiment results show that our model is effective in solving the task of mandarin prosodic boundary prediction, in which the overall prediction performance is improved by 0.8%, and the model size is reduced by about 55%.

Index Terms: prosodic boundary prediction, mandarin, Multi-Task Learning, BLSTM-CRF

1. Introduction

Prosody structure plays an important role in naturalness and intelligibility of mandarin speech synthesis. Unlike English, mandarin is a kind of continuous writing language, so the prosody structure of mandarin is more complex than that of English. In general, mandarin prosody structure is defined as a three-level tree structure including Prosodic Word (PW), Prosodic Phrase (PPH) and Intonational Phrase (IPH) [1]. For example, the mandarin sentence "本文主要研究韵律结构的预测 (This paper mainly studies the prediction of prosody structure)", its prosody structure analysis result is shown in Figure 1. The leaf nodes in bottom layer are Chinese Character (CC), several CCs can be combined into Lexicon Word (LW), several LWs can be combined into PW, then PWs to PPH, and PPHs to IPH.

In recent years, with the development of speech synthesis technology, many researchers also carried out relevant researches on mandarin prosodic boundary prediction. Based on previous studies, [2] use abundant information of syntactic features to improve the performance of mandarin prosodic boundary prediction. Zheng make use of joint learning of word embedding and fusion of different word-level models to improve the performance [3]. And then they also employ BLSTM-CRF model based on Chinese character features to predict mandarin prosodic boundaries, and achieved a good result [4]. In brief, previous studies on mandarin prosodic boundary prediction mainly focused on two main aspects:

1) studying the impact of different kinds of features [1,2,4];
2) exploring suitable models and architectures [3-4].

In previous studies, the task is usually decomposed into three independent subtasks: PW, PPH and IPH, which are modeled and processed respectively. This modeling method ignores the dependencies among subtasks, which may degrade the overall performance.

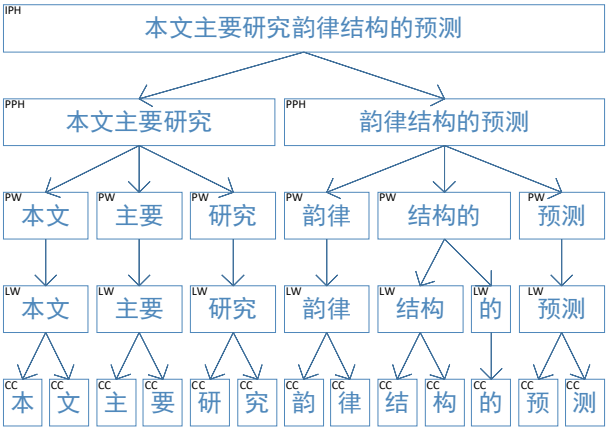

Figure 1: *Example of mandarin prosody structure. (Note: CC-Chinese Character, LW-Lexicon Word, PW-Prosodic Word, PPH-Prosodic Phrase, IPH-Intonational Phrase)*

This paper proposes a mandarin prosodic boundary prediction model based on MTL, which regards PW, PPH and IPH as three subtasks, and uses BLSTM-CRF as the basic model for each subtask respectively. The three subtasks are optimized in a single MTL framework. As shown in Figure 1, general input features of subtasks, such as Chinese character, Chinese word segmentation, part of speech, word length and distance features, could be shared among subtasks by parameters of neural network layers. The relation among subtasks can be represented by adding appropriate connections between the corresponding modules. In training stage, the model is adjusted according to the overall loss of all subtasks, and the global optimal solution could be obtained theoretically.

2. MTL-based mandarin prosodic boundary prediction model

Figure 2 shows the structure of our mandarin prosodic boundary prediction model based on character-level features. There are two parts in the model: 1) feature sharing module, where model

http://dx.doi.org/10.21437/Interspeech.2019-1400

parameters are shared among subtasks; 2) task-specific module, with the connections between input and output of subtasks to reflect the relations among subtasks. For example, both PPH and IPH use the output of the PW subtask as input, so PW will be used as part of the input of the PPH and IPH subtask.

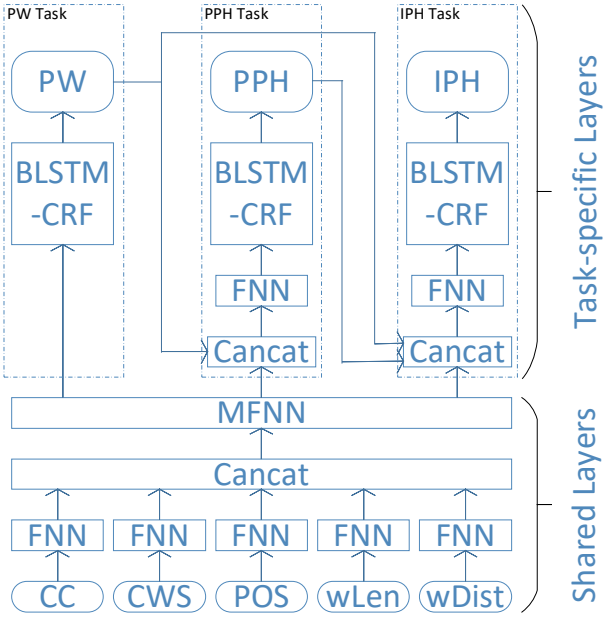

Figure 2: *The general architecture of mandarin prosodic boundary prediction model based on MTL. (Note: CC-Chinese Character, CWS-Chinese Word Segmentation, POS-Part of Speech, wLen-Length of word, wDist-Distance of word, FNN-Feedforward Neural Network, Concat-concatenation, MFNN-Multilayer Feedforward Neural Network)*

2.1. Feature sharing module

In this paper, character-level features are used as input, including Chinese character (CC), Chinese word segmentation (CWS), part of speech (POS), word length (wLen), word distance (wDist) and so on. The feature sharing module transforms each kind of one-hot feature into embedded feature through separate Feedforward Neural Network (FNN) first. For example, assuming the size of dictionary is N_{CC} and embedding dimension is D_{CC}, the embedded feature of CC can be obtained by the following formula:

$$EMB_{CC} = X_{1\times N_{CC}} \times W_{N_{CC}\times D_{CC}} + B_{CC} \quad (1)$$

Among them, $W_{N_{CC}\times D_{CC}}$ is the weight matrix of CC embedding, B_{CC} is the bias term, $X_{1\times N_{CC}}$ is the one-hot feature and EMB_{CC} is the feature embedding. Similarly, the embedded features of CWS, POS, wLen and wDist are calculated as follows:

$$EMB_{CWS} = X_{1\times N_{CWS}} \times W_{N_{CWS}\times D_{CWS}} + B_{CWS} \quad (2)$$

$$EMB_{POS} = X_{1\times N_{POS}} \times W_{N_{POS}\times D_{POS}} + B_{POS} \quad (3)$$

$$EMB_{wLen} = X_{1\times N_{wLen}} \times W_{N_{wLen}\times D_{wLen}} + B_{wLen} \quad (4)$$

$$EMB_{wDist} = X_{1\times N_{wDist}} \times W_{N_{wDist}\times D_{wDist}} + B_{wDist} \quad (5)$$

After feature embedding, a multi-layer Feedforward Neural Network (MFNN) with tanh activation function is added to enhance feature extraction. MFNN sub-network embeds and splices all the features as input and output a whole feature embedding $FEAT_{embed}$ by the following formulas:

$$FEAT_{Concat} = [EMB_{CC}; EMB_{CWS}; EMB_{POS}; EMB_{wLen}; EMB_{wDist}] \quad (6)$$

$$FEAT_{embed} = MFNN_{tanh}(FEAT_{Concat}) \quad (7)$$

2.2. Task-specific module

As shown in Figure 2, we decomposed the mandarin prosodic boundary prediction task into three subtasks, and used BLSTM-CRF as the basic model to deal with each subtask separately. Then, we applied MTL to unify the subtasks under one framework for overall optimization.

For PW subtask, as it does not depend on any other subtasks, so it only takes the whole feature embedding as input. Then we obtain the PW prediction result PW_{pred} through $BLSTM-CRF_{PW}$ sub-network. The formal description is as follow:

$$PW_{pred} = BLSTM-CRF_{PW}(FEAT_{embed}) \quad (8)$$

For PPH subtask, it depends on the PW prediction. So, besides taking the whole feature embedding as input, it also uses PW prediction result. In order to keep the input dimension unchanged, an FNN layer with tanh activation function is added to fuse the whole feature embedding and PW prediction result. The formal description is as follow:

$$PPH_{in} = FNN_{tanh}([FEAT_{embed}; PW_{pred}]) \quad (9)$$

Then through the $BLSTM-CRF_{PPH}$ sub-network, we can get the PPH prediction result PPH_{pred}. The formal description is as follow:

$$PPH_{pred} = BLSTM-CRF_{PPH}(PPH_{in}) \quad (10)$$

For IPH subtask, it depends on the prediction result of PW and PPH subtask both, so we take the whole feature embedding and PW/PPH prediction results as input. Similarly, for the sake of keeping the input dimension unchanged, an FNN layer with tanh activation function is added to fuse the whole feature embedding and the prediction results of PW and PPH. The formal description is as follow:

$$IPH_{in} = FNN_{tanh}([FEAT_{embed}; PW_{pred}; PPH_{pred}]) \quad (11)$$

Then we can obtain the IPH prediction result IPH_{pred} by the $BLSTM-CRF_{IPH}$ sub- network. The formal description is as follow:

$$IPH_{pred} = BLSTM-CRF_{IPH}(IPH_{in}) \quad (12)$$

In accordance with the above treatment, we can get three prediction results of PW, PPH and IPH at one time, and then the complete result of prosody prediction can be obtained by simple merging processing. The label merging process is the reverse process of the decomposition process, and the specific operations can be referred to Section 3.2.

2.3. Loss function

We take the mandarin prosodic boundary prediction as a whole task, so we merge the losses of all subtasks into a whole loss. The loss of each subtask is formally described as follows:

$$Loss_{PW} = Cost_{PW}(PW_{pred}, PW_{ans}) \quad (13)$$

$$Loss_{PPH} = Cost_{PPH}(PPH_{pred}, PPH_{ans}) \quad (14)$$

$$Loss_{IPH} = Cost_{IPH}(IPH_{pred}, IPH_{ans}) \quad (15)$$

In which, $Cost_*$ is loss function of subtasks, PW_{ans}, PPH_{ans} and IPH_{ans} are answer label sequence of subtasks. Then, we can obtain the whole loss by weighted summation of the losses

of each subtask. Assuming that W_{PW}, W_{PPH} and W_{IPH} are weight coefficients of subtask losses of PW, PPH and IPH respectively, the overall loss can be obtained as follow:

$$Loss_{total} = W_{PW} * Loss_{PW} + W_{PPH} * Loss_{PPH} + W_{IPH} * Loss_{IPH} \quad (16)$$

Among them, the weight coefficient of subtasks can be set according to experience, and also can be obtained by other methods such as grid search.

3. Experiment

In order to verify the effectiveness of the proposed method, we take BLSTM-CRF as the benchmark model, and construct three mandarin prosodic boundary prediction models based on character-level features for PW, PPH and IPH respectively. The construct of MTL model refers to section 2.

3.1. Dataset and evaluation metrics

As there is no public dataset for mandarin prosodic boundary prediction task, the experimental data used in this paper is from Databaker[1], which is labelled by two linguistic experts with rich experience. The tagging results have been double-checked to ensure the consistency and accuracy. The dataset contains about 150,000 sentences, which is divided into training set and test set with ratio 9:1. The statistical information of experimental data is shown in Table 1.

Table 1: *Basic Statistical information of dataset.*

Type	Training set	Test set
Sentence	134,997	15,000
CC	2,693,395	299,061
PW	731,074	81,254
PPH	192,668	21,416
IPH	231,774	25,807

The experimental results were evaluated comprehensively by accuracy (ACC), recall (REC) and F1 value (F1).

3.2. Data preprocessing

The data pretreatment process is introduced through the same example sentence in section 1. As shown in Figure 3, data pretreatment mainly includes the following four steps:

1) recovering the labeled prosody sample into original sentence;
2) obtaining CWS and POS information of original sentence through related tools;
3) getting word length and distance information based on CWS and POS results from step 2;
4) aligning all information extracted in the first three steps to CC-level. There is some special processing involved in the alignment: the CWS information needs to be converted into 'BMES' labels, and word length information needs to be transformed into position index of CC in the word by two directions, and word distance information needs to be converted by a similar measure for word length.

All CC-level input features can be obtained through the above four steps. The word segmentation and part-of-speech tagging tools used in this paper come from an existing internal toolkit which is based on CRF model.

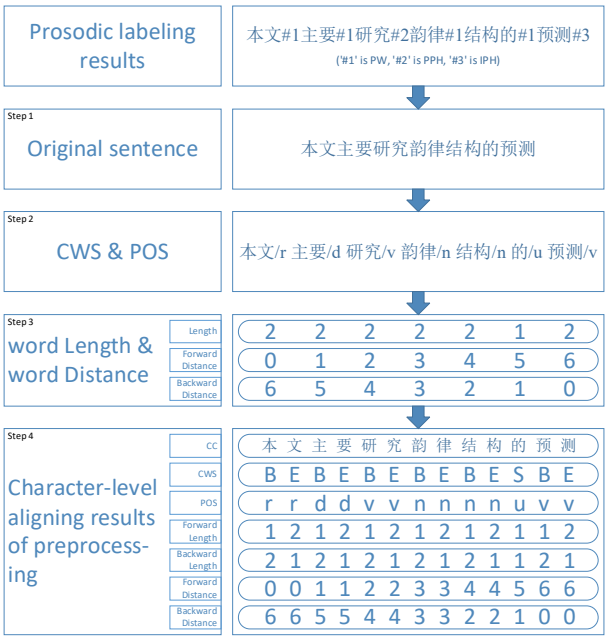

Figure 3: *Example of Chinese character-level features extraction process. (Note: CWS-Chinese Word Segmentation, POS-Part of Speech, word Distance-Distance of current word to head and tail of sentence)*

When the input Chinese character-level features are ready, we process the output sequence. First, we can easily extract mixed label sequence from labeled prosody sample. Then, we can obtain single label sequence of PW, PPH and IPH based on mixed label sequence. Among them, PPH and IPH labels need to be degraded to PW labels in PW single label sequence, and IPH labels need to be degraded to PPH labels in PPH single label sequence. The conversion process is described in Figure 4.

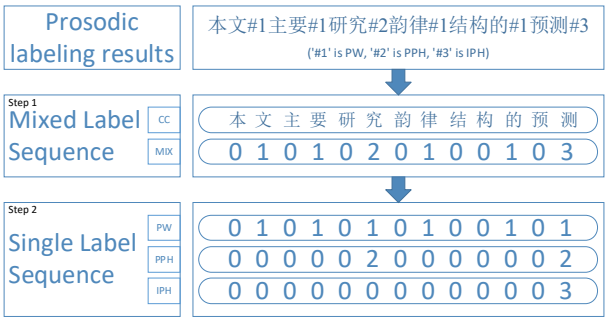

Figure 4: *Example of prosodic label sequence process.*

[1] http://www.data-baker.com/index_en.html

3.3. Contrast experiment I

This group of comparative experiments compare the performance between BLSTM-CRF based single task models and MTL model. The key hyperparameters of the four models are set by the same value, such as batch size, embedding dimension, number of hidden units and layers about BLSTM and so on. In addition, the loss weight ratio of PW, PPH and IPH in MTL model is set to 1.0:1.0:1.0. The experimental results are shown in Table 2.

Table 2: *Experimental results of benchmark models and MTL model.*

Model	Label	ACC	REC	F1
BLSTM-CRF	PW	96.85%	98.06%	**97.45%**
	PPH	84.41%	82.40%	83.40%
	IPH	88.99%	88.51%	88.74%
MTL	PW	96.63%	98.24%	97.43%↓
	PPH	83.91%	83.86%	**83.89%**↑
	IPH	88.87%	89.30%	**89.09%**↑

Compared with the benchmark model, F1 value of MTL model on PPH and IPH task is improved by 0.49% and 0.35% respectively, and the F1 value of PW task has a slight decrease of 0.02%.

Table 3: S*ize of benchmark models and MTL model.*

Model	Label	Model Size
BLSTM-CRF	PW	5.3MB
	PPH	5.3MB
	IPH	5.3MB
MTL	PW PPH IPH	7.1MB

About the model size, which shown in Table 3, the total size of MTL model is about 45% of benchmark model, and is more compact. Benefited from sharing some model parameters in MTL model, the size of MTL model is significantly reduced when compared with benchmark models. That is of great benefit for engineering (memory occupancy, CPU resource consumption, etc.). Therefore, in a practical point of view, MTL model will be a better choice for product and service on the premise that performance indicators are comparable.

3.4. Contrast experiment II

We fixed other hyperparameters, and compared the prediction performance under different loss weight ratios of PW, PPH and IPH in MTL model. According to our experience, the order of importance about mandarin prosodic labels in mandarin speech synthesis is: PW < PPH < IPH, so we set loss weight ratio to 0.5:1.0:2.0. The experimental results are shown in Table 4.

From the result, we find that, with the emphasis on IPH loss and the decrease on PW loss, the F1 value of IPH task increases about 0.06%, while that of PW and PH task decrease 0.11% and 0.18% respectively. Relatively, the prediction performance under loss weight ratio of 0.5:1.0:2.0 is lower than expected.

Table 4: *Experimental results of MTL models under different loss weight ratios.*

Loss Weight Ratio	Label	ACC	REC	F1
1.0:1.0:1.0	PW	96.63%	98.24%	**97.43%**
	PPH	83.91%	83.86%	**83.89%**
	IPH	88.87%	89.30%	89.09%
0.5:1.0:2.0	PW	96.68%	97.98%	97.32%↓
	PPH	84.54%	82.90%	83.71%↓
	IPH	89.25%	89.04%	**89.15%**↑

4. Conclusions

In this paper, we propose an MTL-based mandarin prosodic boundary prediction model. Based on the BLSTM-CRF model, PW, PPH and IPH subtasks are modeled respectively. And then, MTL is used to integrate the three subtasks into one framework for overall optimization. From the perspective of model performance and model size, MTL model is better than BLSTM-CRF based single task model, and more suitable for engineering applications. In the future, we will explore the following aspects: 1) adding pre-training word vector as input; 2) optimizing the method of adjusting the loss weight ratio of PW, PPH and IPH. Some scholars have done some researches on this problem recently [8-9]. In the future, this kind of method will be used to optimize the performance of mandarin prosodic boundary prediction further.

5. References

[1] M. Chu and Y. Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," *Computational Linguistics and Chinese Language Processing,* vol. 6, no. 1, pp. 61-82, 2001.

[2] H. Che, J. Tao and Y. Li, "Improving Mandarin Prosodic Boundary Prediction with Rich Syntactic Features," in *INTERSPEECH*, 2014, pp. 46–50.

[3] Y. Zheng, Y. Li, Z. Wen, X. Ding, and J. Tao, "Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach," in *INTERSPEECH*, 2016, pp. 3201-3205.

[4] 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.

[5] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *international conference on machine learning*, 2008, pp. 160-167.

[6] K. Hashimoto, C. Xiong, Y. Tsuruoka and R. Socher, "A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks," in *empirical methods in natural language processing*, 2017, pp. 1923-1933.

[7] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv preprint arXiv:1706.05098*, 2017.

[8] O. Sener and V. Koltun, "Multi-Task Learning as Multi-Objective Optimization," in *neural information processing systems*, 2018, pp. 525-536.

[9] R. Cipolla, Y. Gal and A. Kendall, "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *computer vision and pattern recognition*, 2018, pp. 7482-7491.