INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

A Mask-based Model for Mandarin Chinese Polyphone Disambiguation

Haiteng Zhang, Huashan Pan, Xiulin Li

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

Abstract

Polyphone disambiguation serves as an essential part of Mandarin text-to-speech (TTS) system. However, conventional system modelling the entire Pinyin set causes the case that prediction belongs to the unrelated polyphonic character instead of the current input one, which has negative impacts on TTS performance. To address this issue, we introduce a mask-based model for polyphone disambiguation. The model takes a mask vector extracted from the context as an extra input. In our model, the mask vector not only acts as a weighting factor in Weighted-softmax to prevent the case of mis-prediction but also eliminates the contribution of non-candidate set to the overall loss. Moreover, to mitigate the uneven distribution of pronunciation, we introduce a new loss called Modified Focal Loss. The experimental result shows the effectiveness of the proposed mask-based model. We also empirically studied the impact of Weighted-softmax and Modified Focal Loss. It was found that Weighted-softmax can effectively prevent the model from predicting outside the candidate set. Besides, Modified Focal Loss can reduce the adverse impacts of the uneven distribution of pronunciation.

Index Terms: polyphone disambiguation, mask vector, Weighted-softmax, Modified Focal Loss

1. Introduction

Mandarin G2P (Grapheme-to-phoneme) module serves to predict corresponding Pinyin sequence for characters, which consists of polyphone disambiguation, tonal modification and retroflex suffixation [1], etc. Polyphone disambiguation, aiming to predict the correct pronunciation of the given polyphonic characters, is an essential component of Mandarin G2P conversion system. According to the research [1, 2, 3, 4], the difficulty of Mandarin polyphone disambiguation mainly lies in heteronyms. Their pronunciations cannot be determined simply by the word itself but require more lexical information and contextual information，such as Chinese word segmentation, POS (part of speech) tagging, syntactic parsing and semantics.

The earliest approaches of polyphone disambiguation mainly relied on dictionary and rules. The pronunciations of polyphonic characters were decided by a well-designed dictionary and some rules crafted by linguistic experts [1, 2]. However, the rule-based method requires a massive investment of labor to build and maintain a robust dictionary. As the amount of data increased, statistical methods were later widely applied in polyphone disambiguation. Experimental results have confirmed the competency of the statistical methods such as Decision trees (DT), Maximum Entropy (ME) to achieve reasonable performance [3, 4]. However, statistical approaches also ask for considerable effort for feature engineering.

The Recent tremendous success of the neural network in various fields has driven polyphone disambiguation to turn to neural network-based models. [5] addressed the task as sequence labelling and adopted bidirectional long-short-term memory (BSLTM) architecture to predict the pronunciation of the input polyphonic characters, which proved that the BLSTM could benefit the task. [6] combined multi-granularity features as input and yielded improvement on the task. The recent emergence of pre-trained model [7-11] made researchers set out to look at polyphone disambiguation based on these models. With the powerful semantic representation, the pre-trained model helps the system to achieve better performance. Bidirectional encoder representations from Transformer (BERT) was applied in front-end of Mandarin TTS system and showed that the pre-trained model outperforms previous methods [12]. Transformer based neural machine translation (NMT) encoder also has a positive effect on the task [13]. However, to avoid the case of prediction belongs to the unrelated polyphonic character rather than the current input one, it is either to model each polyphonic character separately or to uniformly model the entire Pinyin set but adding limitation in the output layer. Yet, the drawback of the former is complex maintenance due to its large number of models, while the latter only limits the prediction output but ignores the impact of the restriction on other modules in the training process. Besides, the unbalanced distribution among polyphones also harmful to the task.

To address these issues, we propose a mask-based architecture for Mandarin polyphone disambiguation by employing a mask vector. In the proposed framework, features including mask vector are taken as input. Then, we apply an encoding layer composed of BLSTM and convolutional neural network (CNN) to obtain semantic features. The Weighted-softmax is latter used to pick up the pronunciation for the polysyllabic character. In the proposed model, the roles that mask vector plays can be concluded as follows 1) Mask vector enriches the input features. 2) Mask vector acts as a weighting factor in Weighted-softmax to prevent the model from mis-predicting the Pinyin of other polyphonic characters. 3) Constraints of candidates by mask vector will pass to the calculation of loss function then better guide the training process. In this way, the proposed approach not only can model the entire polyphonic characters set within one model but also eliminates the case of mis-prediction without harming the training process. Specifically, to mitigate the uneven distribution of pronunciation among polyphonic characters, we introduce a new loss function called Modified Focal Loss. Our experiments demonstrate that the proposed approach can avoid predicting outside the candidate set and ease the imbalanced distribution without harming the performance.

The organization of this paper is listed as follow. Section 1 reviews the background of polyphone disambiguation. Section 2 introduces various input features of our model. Section 3 briefs on our model structure. Section 4 presents the experimental details and results of this thesis. Section 5 gives the conclusions and looks to future research prospects.

http://dx.doi.org/10.21437/Interspeech.2020-1142

2. Features

According to research [1-4], features such as Chinese word segmentation, POS tagging, and contextual information are essential to the task. Therefore, we apply the Chinese characters of the input sentence and the corresponding lexical information, such as Chinese word segmentation and POS tagging, as input features. As we assume only part of characters needs to be disambiguated in a sentence, we utilize a flag token to identify whether the current input character is disambiguation-needed or not. Meanwhile, the polysyllabic characters, apart from Chinese characters features, also have an extra feature to enhance the information provided. Additionally, we adopt a mask vector to restrict the relationship between polyphonic characters and their relevant Pinyin candidates set. The mask vector which consists of boolean value denotes the related Pinyin of the input polyphonic character. For instance, the polyphonic character "会" can be pronounced as "hui4" and "kuai4", and the corresponding pronunciations in mask vector would be assigned a value 1 while the other pronunciations in mask vector would be assigned a value 0 respectively. We add two additional tokens in the mask vector to indicate the monophonic character and unlabelled polyphonic characters. The mask vector here enriches the input features. Besides, it also acts as a weighting factor in Weighted-softmax, which will be described in Session 3.

Finally, we convert the characters sequence to embedding as the model input, along with the auxiliary features mentioned above from the corresponding sentence.

In summary, the proposed model uses a total of six features, including Chinese character, Chinese word segmentation, POS tagging, polyphones, flag token, and mask vector. Details of the various features are described below:

- **Chinese Character (CC)**: Character including monophonic characters and polyphonic characters;
- **Chinese Word Segmentation (CWS)**: Word segmentation results at the character level, which are represented by {B, M, E, S} tags;
- **Part of Speech (POS)**: We perform POS tagging toward input sentence and assign the tag into character level;
- **Polyphones (PP):** A collection of all polyphonic characters within the corpus along with a non-polyphone token;
- **Flag Token (Flag):** The value range is {0, 1, 2}. Each respectively denotes current char that is disambiguation-needed, disambiguation-needless, and monosyllable;
- **Mask Vector (Mask):** The dimension of mask vector equals to the length of the Pinyin set plus with two special tokens "<UN_LABEL>" and "<NO_LABEL>". The former token denotes monophonic characters while the latter one denotes the polyphonic characters that do not require disambiguation;

In the example sentence "仅会在行业规范和会计制度方面进行指导"(It will only provide guidance in occupational standards and accounting system.), we assumed only a part of polyphonic characters in the sentence would be labelled: The first "会" (target candidate set is [hui4, kuai4], the correct pronunciation is hui4), "行" (target candidate set is [hang2, xing2], the correct pronunciation is hang2) and "和" (target candidate set is [he2, he4, huo4, huo2, hu2], the correct pronunciation is he2); and other polyphonic characters such as the second "会" and the second "行" would not be labelled. Relevant input features of the example sentence are shown in Figure 1.

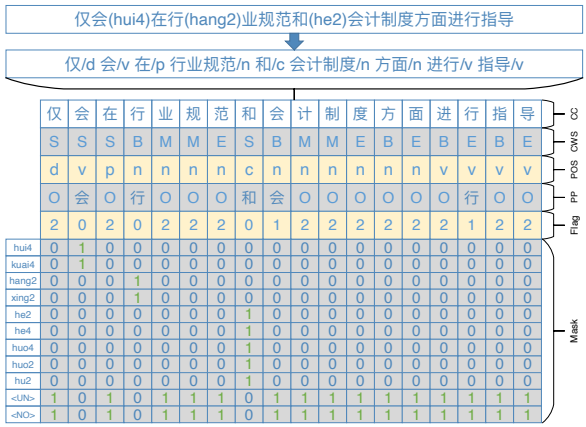

Figure 1: *Input features of the given sample sentence.*

3. Mask-based Mandarin Chinese polyphone disambiguation model

Figure 2 depicts the proposed model's architecture which is mainly composed of three parts as below:

1. **Character-level Feature Embedding Layer**:

   This layer serves to integrate various input features accompanying the mask vector into a low-dimensional and dense vector. First, multiple features are converted into a one-hot label respectively that will be later transformed into an embedding vector by FNN (Feedforward Neural Network). Then, different features' embeddings are concatenated and transformed into a fixed-length vector by MFNN (Multi-layers Feedforward Neural Network).

2. **Context Features Encoding Layer**:

   Accepting a sequence of vectors from the character-level feature embedding layer, this module first extracts semantic information of sentence by both BLSTM and 1D-CNN. FNN layers then intergrade obtained context sequence into a dense vector to represent each word inside the sentence. The reason that motivates us to utilize both BLSTM and 1D-CNN to jointly encode contextual information is mainly based on the following considerations: 1) The BLSTM has an elegant way of encoding sentence-level information. This is extremely helpful when it comes to tasks that need long-range context. 2) 1D-CNN is effective in extracting n-grams level contextual features that are critical for the task of polyphone disambiguation [14, 15].

3. **Restricted Output Layer**:

   To restrict the target candidate set of the current input polyphonic character, the restricted output layer applies the Weighted-softmax by combining the mask vector with softmax to pick up the highest probability within the candidate set. In addition, the proposed model adopts Modified Focal Loss rather than cross-entropy as loss function.

In this work, we explore the Weighted-softmax and Modified Focal Loss modules in terms of improving the performance of polyphone disambiguation.

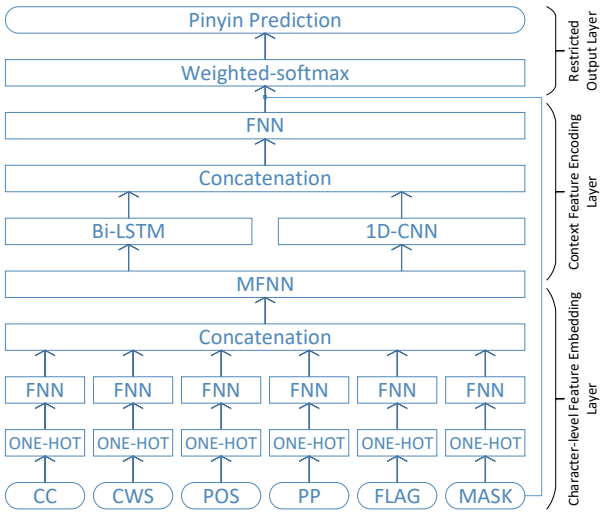

Figure 2: *Network architecture of the proposed model.*

3.1. Weighted-softmax

For each polyphonic character, its candidate range of pronunciation is very limited, which only occupies a small part of the entire Pinyin set. According to softmax, each Pinyin would assign a non-zero probability, leading to the sum of the probabilities obtained in the candidate set is less than 1. It would arouse additional errors, thus making the overall loss larger, which would in turn produce a negative influence on the training process. To address this issue, we constructed the Weighted-softmax by regarding the mask vector as the weighting factor in softmax.

Supposed the input vector of softmax is $V = \{v_1, v_2, \ldots, v_n\}$ and v_i represent the i^{th} element of vector V. For Weighted-softmax, the probability of each element is implemented as follows:

$$p_i = \frac{m_i * e^{v_i}}{\sum_{j=1}^{n} m_j * e^{v_j}} \quad (1)$$

where mask vector is denoted as $M = \{m_1, m_2, \ldots, m_n\}$ and m_i is a Boolean value to denote whether to mask element v_i. By Weighted-softmax, we can assure that the probability of non-candidate pronunciation will not be allocated, and the sum of the probability assigned by the candidate pronunciation set is equal to 1. In this way, we can effectively prevent the model from predicting Pinyin outside the candidate set. Moreover, when calculating loss, Weighted-softmax eliminates the influence of the non-candidate Pinyin set brought to models, thus focusing on the candidate Pinyin set.

3.2. Modified Focal Loss

Due to the uneven distribution of Pinyin, attaching excessive attention to massive and easily classified examples makes the model less precise in terms of rare and hard classified examples, thereby degrading the performance of the system. Concerning uneven distribution in pronunciation among polyphones, inspired by [16], we introduce a novel loss named Modified Focal Loss (MFL) by adding a tunable confidence parameter α to Focal Loss. In Modified Focal Loss, α serves as a threshold to distinguish between massive/easy examples and rare/hard examples, thereby down-weighting the contribution of the former one and up-weighting that of the latter. In this way, Modified Focal Loss enables the model to better classify rare and hard examples.

The equation of Focal Loss is as follows:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t) \quad (2)$$

where p_t denotes the model's predicted probability for the true label and value range is $[0, 1]$. We propose to add the confidence parameter to Focal Loss and the proposed Modified Focal loss is defined as follows:

$$MFL(p_t) = -(1 + \alpha - p_t)^{\gamma} \log(p_t) \quad (3)$$

both α and γ are hyper-parameters, α denotes the tunable confidence parameter and value range is $(0.0, 1.0)$; γ denotes the tunable focusing parameter and value range is $(0, +\infty)$. When the system's estimated probability for the true pronunciation is greater than α, the current input polyphonic character is considered to be easy to classify, and the loss of the corresponding sample will be down-weighted to the overall loss. On the contrary, the input polyphonic character is considered to be difficult to classify, and its loss to the overall loss will be enhanced.

4. Experiment

4.1. Dataset

To verify the proposed method, the experiments were conducted on the dataset from DataBaker[1]. In the corpus, there are 692,357 sentences, and each one at least contains one polyphonic character. We split the dataset into a training set with 623,320 sentences and a test set with 69,037 sentences. Table 1 illustrates the statistical information of corpus.

Table 1: *Statistical information of corpus.*

Character	Polyphone	Training set	Test set
量	liang4	5,402	571
	liang2	156	20
当	dang1	9,070	1,060
	dang4	720	80
相	xiang1	6,003	659
	xiang4	785	74
.....			
Overall	-	623,320	69,037

As in table 1, the frequency of different pronunciations inside a polyphonic character varies greatly both in the training set and test set. The polyphonic characters "量" can be pronounced as "liang4" and "liang2". However, Pinyin 'liang4' appears 5,402 times in the training set and 571 times in test set which are much larger than that of 'liang2'. The same situation occurs when it comes to polyphonic characters "当" and "相". This reveals the uneven distribution in polyphones within dataset.

4.2. Experimental Setting

We implemented the following five systems and used accuracy rate as evaluation criteria for comparing:

1. **BLSTM**: Strictly following the description in [5], we implemented BLSTM model for the task as baseline. NLPIR is adopted for Chinese word segmentation and

[1]https://www.data-baker.com/bz_dyz_en.html.

POS tagging on the input sentence. We set the layers of BLSTM to 2 and the hidden size to 512. The contextual size of polyphonic characters is set to 1 to construct the POS sequence according to [5].

2. **B-CNN**: The input sequence included Chinese character, Chinese word segment, POS tagging, polyphones token and flag token. Context features encoding layer that consists of BLSTM and CNN was to capture the long-range context features. The number of layers in BLSTM and CNN were both set to 2. The hidden size of BLSTM was set to 512. The setting of strides of CNN were 2,3,4, and the kernel num was 64. Rather than modelling the context word, we treated the input sentence as an instance to model. In the training process, we adopted the Adam as optimizer and set the learning rate to 5.0e-4. We split the corpus into mini batch with the batch size of 128.
3. **BC-W**: Same as system 2 but applied Weighted-softmax in the model additionally.
4. **BC-F**: Same as system 3 but applied Focal Loss in the model. The parameter γ was set to 0.7.
5. **BC-WM**: Same as system 3 but applied Modified Focal Loss in the model additionally. The parameter α was set to 0.5, and the parameter γ was set to 0.7.

4.3. Results and analysis

4.3.1. Evaluation of different systems

Table 2 reveals the accuracy of polyphone disambiguation in different systems. It can be seen that B-CNN outperformed BLSTM baseline model. Besides, BC-W gained a better result than B-CNN, verifying the feasibility of the mask vector to strengthen the input features and allow the model to focus on the candidate polyphone. BC-F got similar performance as BC-W. Particularly, BC-WM method achieved the best performance, showing that the Modified Focal Loss can alleviate the imbalance of polyphone distribution.

Table 2: *The accuracy for different system.*

System	BLSTM	B-CNN	BC-W	BC-F	BC-WM
Acc	95.55	97.44	97.85	97.82	**97.92**

4.3.2. Impact of Weighted-softmax

To illustrate the impact of Weighted-softmax, in the case of "他提醒大家明天依旧要注意防晒防中暑" (He reminds everyone to protect from the sun and avoid heatstroke tomorrow), we draw the estimated probability distribution of the polyphonic character "中" on the part of polyphones in Figure 3. As shown, the darker of the location, the greater probability of corresponding pronunciation. Figure 3(a) displays the probability distribution of the prediction towards "中" from the system B-CNN, while Figure 3(b) represents the probability distribution from the system BC-W. In this sentence, the true pronunciation of the polyphonic character "中" is "zhong4 ". As shown in Figure 3(a), the B-CNN system predicts the pronunciation to "huan2" which is not reasonable. Moreover, the probability was allocated to the entire Pinyin set rather than the candidate Pinyin set. As for system BC-WM, only the probability of "zhong1" and "zhong4" are not equal to zero. Both of them are the candidates of "中". Figure 3(b) represents the probability distribution

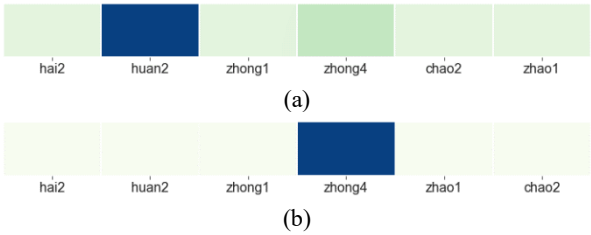

Figure 3: *Probability distribution of* "中".

of prediction from system BC-WM, and the system correctly predicted pronunciation as "zhong4".

4.3.3. Impact of Modified Focal Loss

To illustrate the role of Modified Focal Loss, we collected the accuracy of several polyphones suffered from imbalanced distribution mentioned in chapter 4.1. The accuracy rate from the system BC-W, BC-F and BC-WM are listed as table3.

Table 3: *The accuracy of polyphonic characters.*

Character	Polyphone	BC-W	BC-F	BC-WM
量	liang4	98.60	**99.82**	99.65
	liang2	**70.00**	55.00	**70.00**
当	dang1	98.49	98.77	**99.39**
	dang4	80.00	77.50	**87.50**
相	xiang1	99.24	98.94	**99.39**
	xiang4	94.59	94.59	**95.95**

The experimental results revealed that Modified Focal Loss is highly conductive to minimize the adverse influence of imbalanced distribution within Pinyin set. The accuracy of "dang4" in BC-WM is 7.5% higher than that of BC-W and 10% higher than that of BC-F. As the case of 'xiang4', BC-WM is 1.36 % higher than that of system BC-W and BC-F. Besides, system BC-WM got a slightly improvement compared to the other systems on massive examples such as "dang1" and "xiang1". It indicates that the Modified Focal Loss can improve competency of the model in classifying rare and hard examples without harming that of the massive examples.

5. Conclusions

In this paper, we proposed a mask-based architecture for Chinese Mandarin polyphone disambiguation, where mask vector is not only a part of input features but also a weighting factor in Weighted-softmax. Besides, we optimized the loss function from cross-entropy to Modified Focal Loss. The proposed architecture can achieve an accuracy rate at 97.92%, a 2.37% improvement compared with that of the baseline model. The experimental results demonstrate that the mask vector can effectively prevent model from predicting outside the candidate set. In addition, Modified Focal Loss can ease the distribution imbalance of Pinyin set.

In the future, we will make the proposed Weighted-softmax and Modified Focal Loss collaborate with pre-trained models such as Elmo and Bert to fulfill the task of polyphone disambiguation.

6. Acknowledgement

The authors would like to thank the data team for their assistance.

7. References

[1] L. Yi, L. Jian, H. Jie, and Z. Xiong, “Improved Grapheme-to-Phoneme Conversion for Mandarin TTS,” *Tsinghua Science &Technology*, vol. 14, no. 5, pp. 606–611, 2009.

[2] H. Dong, J. Tao, a nd B. Xu, “Grapheme-to-Phoneme Conversion in Chinese TTS System,” in *2004 International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 165–168, 2004.

[3] H. Zhang, J. Yu, W. Zhan, and S. Yu, “Disambiguation of Chinese Polyphonic Characters,” in *The First International Workshop on MultiMedia Annotation (MMA2001)*, vol. 1, pp. 30–1, 2001.

[4] 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 (ICGEC)*, pp. 244–247, 2010.

[5] C. Shan, L. Xie, and K. Yao, “A bi-directional lstm approach for polyphone disambiguation in mandarin chinese,” in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, pp. 1–5, 2016.

[6] Z. Cai, Y. Yang, C. Zhang, X. Qin, and M. Li, “Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features,” *Proceedings of Interspeech 2019*, pp. 2110-2114, 2019.

[7] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *Proceeding of the ICLR, 2015*.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[9] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics, (NAACL)* pp. 2227-2237, 2018.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,”. *Proceeding of the NAACL-HLT*, vol.1, pp.4171-4186, 2019.

[11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *Proceeding of the NeurIPS*, vol. 32, pp. 5754-5764, 2019.

[12] D. Dai, Z. Wu, S. Kang, X. Wu, J. Jia, D. Su, and H. Meng, "Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-Trained BERT," *Proceedings of the Interspeech 2019,* pp. 2090-2094, 2019.

[13] B. Yang, J. Zhong, and S. Liu, "Pre-Trained Text Representations for Improving Front-End Text Processing in Mandarin Text-to-Speech Synthesis," *Proceedings of the Interspeech 2019,* pp. 4480-4484, 2019.

[14] Kim Y. “Convolutional neural networks for sentence classification” *Procedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp1746-1751,2014.

[15] J.R.Novak, N.Minematsu, and K.Hirose, “Failure transitions for joint n-gram models and g2p conversion”. In INTERSPEECH, 2013, pp. 1821-1825.

[16] T. Lin, P. Goyal, R. Girshich, K. He, and P Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980-2988, 2017.