2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) | 978-1-7281-8535-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/IMCEC51613.2021.9482008

A Method of Polyphone Disambiguation Based on Semantic Extension

Qi Wang

China University of Geosciences, Beijing, School of Information Engineering
wangqi_cugb@163.com

***Abstract*—In text-to-speech (TTS) system, polyphone disambiguation is the significant and difficult point in the process of grapheme-to-phoneme(G2P) conversion. In the existing methods, polyphone disambiguation mainly uses statistics or context information. When the data to be predicted is short texts or spoken language, it is difficult for the previous models to distinguish the correct pronunciation due to insufficient information. Therefore, this paper intends to report an improved design for polyphone disambiguation. Semantic extension has been used through the method of feature fusion, we creatively encode the context feature, word meaning feature, part of speech feature and pronunciation feature of polyphone, and construct three kinds of polyphone disambiguation models using bidirectional LSTM neural network. It is shown in the experiment that the accuracy of the polyphone disambiguation model based on word meaning can be improved by about 7% compared with the traditional methods.**

Keywords—polyphone disambiguation, feature fusion, semantic extension, bidirectional LSTM

I. INTRODUCTION

In text-to-speech (TTS) system, the grapheme-to-phoneme (G2P) module is an important part. However, there is a situation where one character has more than one pronunciation in Chinese. These words are called polyphone. Thus the pronunciation disambiguation of polyphone is the core and difficult point in the process of text to speech conversion. That means the accuracy of this module directly affects the intelligibility of the TTS system. Polyphone disambiguation is the automatic discrimination of polyphone pronunciation in context information. It is generally believed that the pronunciation of polyphonic words is related to semantics and linguistic habits. For example, Chinese word "还" reads “huan2” when it means "return", and reads "hai2" which means "still". So, it is necessary to understand the meaning of polyphone according to the context semantics and then make pronunciation judgment.

In general, the pronunciation of Chinese polyphone is influenced by the context, the meaning of polyphone and the language habit. Thus, polyphone disambiguation is the automatic method of distinguishing and analyzing the pronunciation of polyphonic words. Existing methods mainly include rule-based or machine-learning-based, deep-learning-based methods. Early polyphone pronunciation disambiguation method is mainly based on the rules or on dictionary. In Dong et al. [1], the disambiguation of polyphone is accomplished by constructing a polyphone dictionary or by using the manual rules to judge the pronunciation of polyphone. Similarly, in the study of Zhang et al.[2], the author constructed a polyphone library to solve the problem of polyphone disambiguation. The library is retrieved by rules to obtain the pronunciation of polyphone when using.

With the increase of the amount of data, more methods use machine learning to solve the task of polyphone disambiguation. For example, in the research of Liu et al.[3], the likelihood ratio and mutual information screening are used to distinguish polyphone, and then the decision tree is used to classify polyphone pronunciation. In the study of Fan et al.[4], the author used the combination of words around polyphone as the feature and the maximum entropy model for pronunciation classification, and achieved better results than the decision tree.

In recent years, with the development of neural network in the field of natural language processing (NLP), the use of neural network for polyphone disambiguation has attracted more and more attention. In the research of Shan et al. [5], the author regarded polyphone disambiguation task as a sequence labeling task, and used bidirectional LSTM neural network to predict polyphone pronunciation. The results are better than the maximum entropy model. The author used the words around the polyphone to predict the pronunciation of polyphonic words by constructing the combination of words and parts of speech. In the research of Cai et al.[6], the author used the multi-layer word embedding feature as the representation of polyphone. He combined the word vector of polyphone and the sentence vector of the text and predicted the pronunciation of a polyphone by multi classification. In the study of Dai et al.[7], the author used Bert[8],a pre-training model with bidirectional LSTM[9] or Transformer[10] to encode the text which contains polyphone, and used the hidden layer of polyphone to

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 21,2022 at 09:33:46 UTC from IEEE Xplore. Restrictions apply.

predict the pronunciation. The results showed that Bert model combined with bidirectional LSTM could achieve the best effect.

The existing polyphone disambiguation mainly uses the above three methods. However, these methods have their own disadvantages. The first is that the rule-based method requires a wealth of knowledge of the language and a robust dictionary to cope with various situations. Easy to cause a rule "disaster". In practice, there are also problems such as incomplete coverage of rule dictionary or low accuracy. And for spoken language data, the above problems are more obvious due to sparse environment text.

For machine learning or deep learning methods, models are less effective when text is short or affected by word segmentation.

Ⅰ、霍天鹰一脸鄙夷地冲李信斥道	→ chong1
Ⅱ、包括东京都没有禁止...	→ du1
Ⅲ、霍天鹰一脸鄙夷地#冲#李信斥道[SEP]冲[SEP]对着，向着。	→ chong4
Ⅳ、包括东京#都#没有禁止[SEP]都[SEP]全，完全，都要。	→ dou1

Fig. 1. A sample polyphone text

As is shown in the Fig.1. above, the texts to be tested are all from online novels. The first sentence contains the polyphonic character "冲", which has two pronunciations of “chong4”and “chong2”. Due to the relatively short ambient text of polyphonic words and the inclusion of names, it is difficult to predict the correct pronunciations with the multi-classification model or manual rules based on context information. Similarly, in the second sentence, the polyphonic word "都" contains the pronunciations "du1" and "dou1". Because of the wrong word segmentation, "…东京都|没有|…" is incorrectly linked and at this time “东京都” means “Tokyo” and is read “du1”. However, the correct word segmentation is “…东京|都没有|…” and the word “都” should be read “du1”. By extending definition of polyphonic words, the model can learn the correct definition and pronunciation of polyphonic words at the same time and improve the accuracy of judgment and can effectively solve the above problems

In this paper, the research mainly focuses on how to integrate the features of the context semantic information, the meaning of polyphone, the part of speech information and the pronunciation information of the polyphone, and using the bidirectional LSTM network as the basic classifier to build the polyphone disambiguation model.

To sum up, the main contributions of this paper are as follows:

- The model integrates the meaning of polyphone, part of speech information and pronunciation information of polyphone into the task of polyphone disambiguation.
- The structure of the polyphone disambiguation model has been redesigned and this method transforms the multi classification problem used in the traditional polyphone disambiguation into the binary classification problem. The model gives more accurate recognition results and has better generalization performance.

II. METHODS

The polyphonic disambiguation model used in this article will be discussed in detail in this section. Firstly, the definition of polyphonic word disambiguation task is described, and then model of polyphonic word disambiguation which combines a variety of features is discussed.

A. Task Definition

In the polyphone disambiguation task, the sentence S is composed of a series of words w_i $S = \{w_0, w_1 \ldots . w_n\}$, where n is the sentence length, in sentence S w_i is a polyphonic word, the pronunciation contains $\{t_0, t_1 \ldots . t_k\}$, where k is the number of pronunciations contained by the word w_i, so the task of polyphone disambiguation is to judge the pronunciations of the polyphonic word w_i through the context semantics. For example, the Chinese character "冲" is pronounced "chong1" when used as a verb, such as "冲锋". When used as a adverbs, it is pronounced "chong4" such as "冲着". However, some individual polyphonic words, such as "都", cannot be judged only by the part of speech. Instead, contextual semantics and other features are needed to assist the judgment. Therefore, the task of polyphonic word disambiguation is formally defined.

$$f : S \rightarrow T \tag{1}$$

$$f(t|w_i) \qquad t \in \{t_0, t_1 \ldots . t_k\} \tag{2}$$

Where S is the sentence containing polyphonic words to be detected, T is the target domain of pronunciation, and the polyphone disambiguation task is to predict the pronunciation of t based on features such as polyphone w_i and surrounding context.

B. LSTM

Recurrent neural network (RNN) is an effective method of text sequence encoding, since it can well represent the relationship between text sequences. In order to solve the problem of gradient explosion in RNN network, the long short-term memory (LSTM) structure was proposed. Compared with RNN, LSTM can better solve the problem of long-term dependence. The formula is as follows:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{3}$$

$$f_t = \sigma(W_t[h_{t-1}, x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5}$$

$$\tilde{C}_t = tanh\,(W_c[h_{t-1}, x_t] + b_c) \tag{6}$$

$$h_t = o_t \times tanh\,(C_t) \tag{7}$$

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 21,2022 at 09:33:46 UTC from IEEE Xplore. Restrictions apply.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (8)$$

The input x_t is the eigenvector, σ represents the $sigmod$ activation function, h_{t-1} is the hidden layer state at the previous moment and h_t is the hidden layer state at the current moment. In addition, i_t, f_t, and o_t represent the input gate, the forgotten gate, and the output gate. C_t represents the abstract information, while $\tilde{C}_t$ represents the information retained at the current moment. Respectively, both W and b are training parameters.

C. Sentence Pair Representation

In order to better integrate the lexical meaning information of polyphonic words into the task of polyphone disambiguation, we constructed the input of Text- Meaning sentence pairs as the model of polyphonic word disambiguation, as is shown in Fig.2. Similar to the natural language inference task, the disambiguation task is regarded as the sentence pair classification task. At the same time, in order to further improve the accuracy of the model, weak supervision mechanism and part of speech information are also considered as classification features. The details are as follows:

Text-Meaning sentence pairs: we connect the text sequence containing polyphonic words with the polyphone meanings information and pronunciation information of polyphonic words under the pronunciation, and use [SEP] for segmentation. Using the same method, one of the polyphonic pronunciations is randomly selected, and the wrong word meaning, pronunciations and text sequences are spliced as negative samples, so as to achieve the balance of positive and negative samples.

包括东京都没有禁止[SEP]都[SEP]都：全，完全，都要。表示语气的加重[SEP]dou1	True
包括东京都没有禁止[SEP]都[SEP]都：全，完全，都要。表示语气的加重[SEP]adv[SEP]dou1	True
包括东京#都#没有禁止[SEP]都[SEP]都：全，完全，都要。表示语气的加重[SEP]adv[SEP]dou1	True
包括东京都没有禁止[SEP]都[SEP]都：大都市，都市，都会[SEP]du1	False
包括东京都没有禁止[SEP]都[SEP]都：大都市，都市，都会[SEP]adv[SEP]du1	False
包括东京#都#没有禁止[SEP]都[SEP]都：大都市，都市，都会[SEP]adv[SEP]du1	False

Fig. 2. The first line shows the text-meaning of sentences pairs, which contains the meaning of the context of polyphone, polyphone word and meaning of polyphone, distinguish with [SEP], also the second sentence add the part of speech information, such as "adv" .The third sentence is increased the weak supervision mechanism .The words "#" is used to mark polyphone. The negative sample is shown in the following three lines by connecting the wrong pronunciation and wrong polyphone meaning with the text to be recognized

Text-Meaning-POS sentence pairs: Through the observation of polyphonic pronunciation samples, part of speech information can be used as an important feature of polyphonic pronunciation disambiguation task. Therefore, the part of speech annotation is carried out for each text sequence of polyphonic words through part of speech annotation tool, and the part of speech of polyphonic words is taken as a feature to be spliced and segmented by [SEP].

Weak supervision mechanism: In order to make the model pay more attention to the predicted polyphonic words, this paper uses weak supervision mechanism to label the predicted polyphonic words. Specifically, "#" is used as the mark on both sides of the polyphonic word in the text sequence to indicate that the word is a polyphonic word sample to be predicted. This enables the model to learn the location of the sample to be tested, rather than focusing on the rest of the text in the sequence.

D. Polyphonic Disambiguation Model

The polyphone disambiguation model is shown in Fig.3. The constructed sentence pairs are encoded by using bidirectional LSTM network.

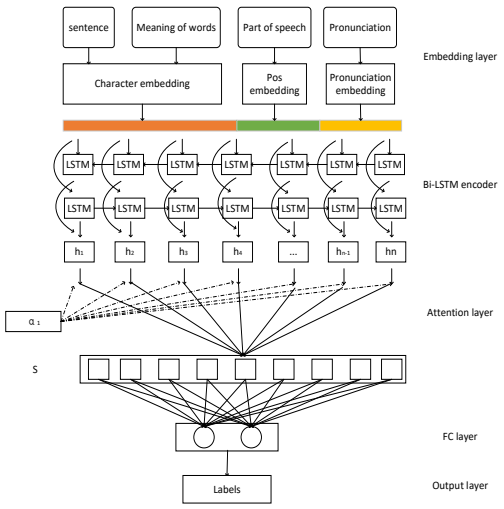

Fig. 3. Polyphone disambiguation model

For an input sequence, $word_i$ is used to represent each word in the sequence, in which $i \epsilon [0, N]$, N is the maximum length of a sentence and different words are mapped into word vectors through word vector matrix, where E_{word} is word vector matrix, e_{word} is the unique expression of each word, and part of speech vector and

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 21,2022 at 09:33:46 UTC from IEEE Xplore. Restrictions apply.

pronunciation vector are similar to word vector. The formula is as follows

$$e_{word} = E_{word} \cdot word_i \tag{9}$$

$$e_{pos} = E_{pos} \cdot pos_i \tag{10}$$

$$e_{pron} = E_{pron} \cdot pron_i \tag{11}$$

$$e_i = concat([e_{word}; e_{pos}; e_{pron}]) \tag{12}$$

The bidirectional LSTM network is used to encode the text, and the forward hidden layer representation and the reverse hidden layer representation are connected as the deep semantic representation of a data. The formula is as follows:

$$h_{fi} = LSTM(e_0, e_1 \dots e_n) \tag{13}$$

$$h_{bi} = LSTM(e_n, e_{n-1} \dots e_0) \tag{14}$$

$$h_i = [h_{fi}; h_{bi}] \tag{15}$$

The important information in the text sequence is learned through the attention mechanism. First the hidden representation h_i is feed to the single-layer MLP to obtain the word representation. Then we measure the importance of the word as by calculating the similarity with the word level context vector u^T, and get the attention weight α_i by $softmax$ function. The attention weight is multiplied by the hidden layer state h_i .The sentence vector s is a high-level representation of the sentence and can be used as features for polyphone disambiguation. Given the sentence representation vector s , we can get category probability p and label y through $softmax$ function and $argmax$ function. The model uses the cross entropy as the training loss.

$$\alpha_i = \frac{\exp(\tanh(W_\alpha h_i + b_\alpha) \cdot u^T)}{\sum_j \exp(\tanh(W_\alpha h_i + b_\alpha) \cdot u^T)} \tag{16}$$

$$s = \sum_{j=0}^{L} \alpha_i h_i \tag{17}$$

$$p = softmax(Relu(W^T s + b)) \tag{18}$$

$$y = argmax(p) \tag{19}$$

III. EXPERIMENTS

A. Data Set

In order to make the model more generalized, we selected a corpus that contains 250000 texts which have polyphonic words, and the text style includes more than 10 categories such as military, agricultural, medical, novel and so on. Most of them are spoken data and come from the Internet. Through manual labeling, each training example contains a polyphonic word to be recognized, and the corpus contains 197 polyphonic words with 397 pronunciations.

TABLEI. THE TOP 5 POLYPHONE

Words	Counts	Proportion
为	36251	14.5%
地	18240	7.2%
得	13952	5.5%
子	12322	4.9%
都	10050	4.0%

TABLEII. THE TOP 5 PRONUNCIATIONS

Phoneme	Counts	Proportion
de5	25765	10.3%
wei2	20527	8.2%
wei4	15732	6.2%
zhong4	7677	3.0%
dang1	6888	2.7%

For the purpose of better describing the data set, we calculated the top five polyphonic words according to word frequency, as can be seen in table 1 and table 2. The first five pronunciations were counted according to their frequencies. It can be seen that the polyphonic word "为" occupies the largest proportion in the data set, followed by "地" "得" "子" "都". Among the pronunciations of polyphonic words, "de5" accounts for the largest proportion, followed by "wei2" and "wei4", which also conforms to the frequency of polyphonic words and polyphonic words in daily life

B. Experiment Setup

First of all, in terms of data sets, we evaluated the effectiveness of our model on our Chinese polyphone dataset. The dataset is divided into training set and test set according to the proportion of 8:2. After data cleaning, the training set contains 198,760 sentences and the test set contains 49,926 sentences. In addition, we use TensorFlow to build a neural network. The size of the hidden layer is 150 dimensions, the size of the word vector is 300 dimensions, and the maximum sentence length is 400.The learning rate was 0.001 and further decreased with training. THULAC[11] tool has been used for part-of-speech tagging and Nvidia Titan XP 12G GPU platform has been used during training.

C. Results

In order to conduct comparative experiments, we constructed several comparative experiments. What's more, we used traditional methods, such as sequence prediction method and Bert polyphonic pronunciation classification method to prove the effectiveness of lexical meaning expansion. More details are as follows:

- **LSTM-Word-Level:** Using LSTM to categorize text sequences containing polyphonic words at word level, and use the final hidden layer representation of a sentence for polyphonic classification.
- **LSTM-Char-Level:** Using LSTM model to encode the sentence with char level, and use the final hidden layer representation of a sentence for polyphonic classification.
- **LSTM-POS:** LSTM is used to encode sentence with word level and to use part of speech as an additional feature for classification of pronunciations.
- **LSTM-Hidden-Level:** Using LSTM as the encoder, the separate Hidden layer at the position of

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 21,2022 at 09:33:46 UTC from IEEE Xplore. Restrictions apply.

polyphonic word is used for polyphonic word pronunciation classification.

- **Bert-Hidden-Level:** Using Bert as the encoder, the separate Hidden layer at the position of polyphonic word is used for polyphonic word pronunciation classification.
- **LSTM-Pronunciation:** Using context information, pronunciation of polyphone as feature for classification.
- **LSTM-Meaning-Pronunciation:** Using context inf ormation, polyphone meanings, pronunciation of pol yphone as feature for classification.
- **LSTM-Meaning-Pronunciation-POS:** Using conte xt information, polyphone meanings, Pronunciation of polyphone with part of speech feature.

The experimental results are shown in the following t able:

TABLEIII. THE EXPERIMENTAL RESULTS

Methods	ACC(%)
LSTM-Word-Level	69.8%
LSTM-Char-Level	85.7%
LSTM-POS	87.55%
LSTM-Hidden-Level	89.12%
Bert-Hidden-Level	93.25%
LSTM-Pronunciation	85.6%
LSTM-Meaning-Pronunciation	94.08%
LSTM-Meaning-Pronunciation-POS	**96.76%**

It can be seen from the TABLEⅢ that the accuracy has been greatly improved after the use of context information, polyphone meaning information and pronunciation information with POS information, and the accuracy has been improved by 3.51% compared with the use of Bert for polyphonic word pronunciation classification. At the same time, compared with the model without polyphone meaning expansion, the accuracy of adding polyphone meaning feature and part of speech feature has been improved, which are 8.48% and 2.68% respectively. It is obvious that the feature of polyphone meaning can effectively improve the accuracy of the speech disambiguation model.

IV. CONCLUSIONS

To sum up, we mainly study how to improve the accuracy of polyphone disambiguation model by using polyphonic word meaning information in this paper. We propose to feature fusion of the context information, word meaning information, part of speech information and pronunciation information of polyphonic word for the judgment of polyphonic word pronunciation. Experiments on the Chinese polyphone dataset constructed by us show that our model performs significantly better than previous methods. And the result illustrates that our model is effective in picking out the correct pronunciation of the polyphone.

REFERENCES

[1] H. Dong, J. Tao, B. Xu, Grapheme-to-phoneme conversion in Chinese TTS system[C], Chinese Spoken Language Processing, 2004 International Symposium on. IEEE, 2005.

[2] L. Zhang, A solution to polyphone problem in Chinese TTS system[J] Computer Applications and Software 2008.

[3] F. Liu, Q. Shi, and J. Tao, Maximum entropy based homograph disambiguation[J], in The 9th National Conference on Man Machine Speech Communication, 2007.

[4] M. Fan, G. Hu, R. Wang, Multi-level Polyphone Disambiguation for Mandarin Grapheme Phoneme Conversion[J]. Computer Engineering and Applications, 2006.

[5] C. Shan, L. Xie, and K. Yao, A bi-directional lstm approach for polyphone disambiguation in mandar in Chinese[J], in 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE,2016.

[6] Z. Cai, Y. Yang, C. Zhang, et al. Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features[C]. conference of the international speech communication association, 2019.

[7] Dai. D, Wu. Z, Kang, S., Wu. X, Jia. J, Su. D, Yu. D, Meng. H, (2019) Disambiguation of Chinese Polyphones in an End-to-End Framework with Semantic Features Extracted by Pre-Trained BERT[J]. Proc. Interspeech 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805,2018.

[9] Hochreiter S, Schmidhuber J. Long Short-term memory[J]. Neural Computation, 1997.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [A]. In Advances in Neural Information Processing Systems [C], 2017.

[11] Z. Li, M. Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation[J]. Computational Linguistics 2009.

Authorized licensed use limited to: BEIJING UNIVERSITY OF TECHNOLOGY. Downloaded on June 21,2022 at 09:33:46 UTC from IEEE Xplore. Restrictions apply.