Phonologically Aware BiLSTM Model for Mongolian Phrase Break Prediction with Attention Mechanism

Rui Liu, FeiLong Bao(✉), Guanglai Gao, Hui Zhang, and Yonghe Wang

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, College of Computer Science, Inner Mongolia University, Hohhot 010021, China
liurui_imu@163.com, csfeilong@imu.edu.cn

Abstract. Phrase break prediction is the first and most important component in increasing naturalness and intelligibility of text-to-speech (TTS) systems. Most works rely on language specific resources, large annotated corpus and feature engineering to perform well. However, phrase break prediction from text for Mongolian speech synthesis is still a great challenge because the data sparse problem due to the scarcity of resources. In this paper, we introduce a Bidirectional Long Short-Term Memory (BiLSTM) model with attention mechanism which uses the position-based enhanced phonological representations, word embeddings and character embeddings to achieve state of the art performance. The position-based enhanced phonological representations, derived from a separately BiLSTM model, are comprised of phoneme and syllable embeddings which take along position information. By using an attention mechanism, the model is able to dynamically decide how much information to use from a word or phonological component. To handle Out-of-Vocabulary (OOV) problem, we incorporated word, phonological and character embeddings together as inputs to the model. Experimental results show the proposed method significantly outperforms the systems which only used the word embeddings by successfully leveraging position-based phonologically information and attention mechanism.

Keywords: Mongolian · Phrase break · Phonologically
Attention mechanism · Position

1 Introduction

Phrase break plays an important role in both naturalness and intelligibility of speech [1]. It breaks long utterances into meaningful units of information and

This research was supports by the China national natural science foundation (No. 61563040, No. 61773224), Inner Mongolian nature science foundation (No. 2016ZD06) and the Enhancing Comprehensive Strength Foundation of Inner Mongolia University (No. 10000-16010109-23).

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11012, pp. 217–231, 2018.
https://doi.org/10.1007/978-3-319-97304-3_17

makes the speech more understandable. Therefore, identifying the break boundaries of prosodic phrases from the given text is crucial in speech synthesis. Many statistical methods have been investigated to model speech prosody, including classification and regression tree [2], hidden Markov model [3], maximum entropy model [4] and conditional random fields (CRF) with linguistic class features (such as part-of-speech (POS), length of word etc.) [5–13]. However, the linguistic class features are discrete linguistic representations of words, which don't take into account the distributional behavior of words. Recent developments in neural architecture and representation learning have opened the door to models that can discover useful features automatically from the unlabelled data. With this development, neural networks and word embeddings have been increasingly investigated in order to minimize the effort of feature engineering and achieved similar or even superior performance over conventional method in phrase break prediction tasks [14–21]. Vadapalli et al. [17] utilized deep neural networks (DNNs) and recurrent neural networks (RNNs) to model phrase break by using word embeddings. Zheng et al. [20] proposed a character-enhanced word embedding model and a multi-prototype character embedding model for Mandarin phrase break prediction. Klimkov et al. [21] investigated how various types of textual features can improve phrase break prediction and BiLSTM and word embeddings proved to be beneficial.

All these methods mentioned have made great contributions, while they heavily rely on the availability of hand-labeled training data. As a result, these methods can not be used for languages where the necessary linguistic resources are not readily available, and manual annotation of data is expensive and time-consuming. Thus it is hard to work with scripts of Mongolian languages, which are agglutinative in nature and lacks sufficient training corpus. A better solution may lie in dealing with sub-word units like stem and suffixes. Liu et al. [22] proposed a suffix segmentation method, they segmented the ending suffix followed by the Narrow No-Break Space (NNBSP) [23–25] in Mongolian nouns according to the characteristics of Mongolia word formation and then treated them as individual tokens to model Mongolian phrase break. However, the naturalness of synthetic speech is less than satisfactory, especially without a good rhythm.

In this work, we make full use of phonological information to model Mongolian phrase break. Our underlying assumption is that prosodic phrases are likely to be transliterated. Additionally, phenomena such as vowel harmony manifest explicitly in phoneme and syllable representation and can potentially be helpful for Mongolian phrase break.

We first identify the sequence of phonemes and syllables automatically and use Bidirectional Long Short-Term Memory (BiLSTM) networks to encode phoneme and syllable level information to a phonological representation, which captures the phonological information of the word. Then we combine word level representation and phonological level representation to a complex representation. In addition, we use character embeddings to handle the Out-of-Vocabulary (OOV) problem as in [26]. At last, the complex representation and character embeddings are comprised together to a joint embedding and we feed it

in another separately BiLSTM to model context information of each Mongolian word and decode the corresponding right phrase break label. Second, we propose a position-based enhanced method for phonological embeddings. General phonological embeddings cannot distinguish between uses of the same phoneme and syllable in different contexts. We add a positional tag, e.g. the first/second/third/etc., for each phoneme and syllable in each word. We learn phoneme and syllable embeddings from each positionally tagged sequence of phonemes and syllables for each word. Third, we also propose an attention module which chooses the most informative component among available ones (in our case, word embeddings, phonological embeddings) to extract context form.

We experiment our approach with Mongolian languages, and the reported results show the proposed approach achieves the best performance. The phoneme and syllable representation provides richer phonological information for word representation and plays an important role in the neural network architecture. Finally, we incorporate the proposed phrase break model in a Mongolian Text-to-Speech (TTS) system and demonstrate its usefulness with listening tests.

Our contributions are the three-fold: (1) We propose a BiLSTM approach to predict Mongolian prosody phrase labels leverage phonologically information from Mongolian phoneme and syllable without any feature engineering. (2) The position-based enhanced phonological embedding method, which takes into account the positional information in different contexts of the same phoneme and syllable, are investigated. (3) We propose a general attention module that selectively chooses information sources to extract primary context form, maximizing information gain from each component: words, phonological information (comprised of phoneme and syllable embeddings).

2 Proposed Model

Figure 1 shows the overall architecture of the Mongolian phrase break prediction model. The set of input features for each token is basically formed by three distinct components: the word embedding (WE), phonologically embedding (PE) derived from phoneme (PhoE) and syllable embeddings (SylE), and character embeddings (CE). For each given token, we first obtain PhoE and SylE, then we concatenate these two embeddings to get a new embedding called PE. Attach the PE to WE to form a complex embedding with attention mechanism. At last, CE is used along with the complex embedding as a joint embedding. We fed the joint embedding into BiLSTM [27,28] phrase break prediction model to decode the corresponding right phrase break label. We formulate each component of the model in the following subsections.

2.1 Input Features

2.1.1 Word Embeddings

Word embedding map words into a space where semantically similar words have similar vector representations [29]. Based on this idea, more and more embedding

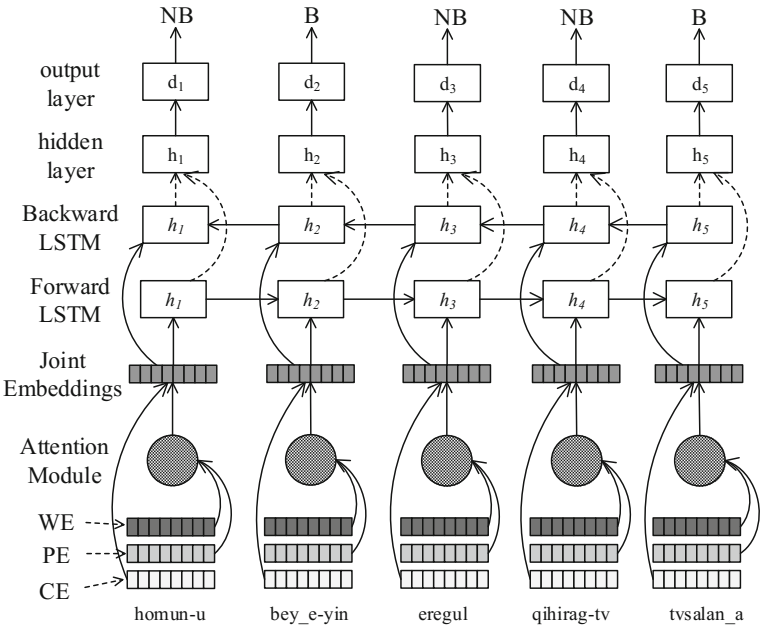

Fig. 1. Mongolian phrase break prediction model. The WE and PE (Fig. 2) are concatenated with attention mechanism, CE is used along with the concatenated embedding to form a joint embedding as input; a BiLSTM produces context-dependent representations; the information is passed through a hidden layer and the output layer. The outputs are either probability distributions for softmax. (WE: word embedding; PE: Phonological embedding; CE: Character embedding)

models have been developed, including continuous bag-of words model (CBOW), Skip-Gram model [30], and Global C&W (GloVe) [31]. We use Skip-Gram model to train the word embedding representation.

The training objective of the Skip-Gram model is to find word representations that are useful for predicting the surrounding words in a raw text. More formally, given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the objective of the Skip-Gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{1}$$

where c is the size of the training context (which can be a function of the center word w_t). Larger c results in more training examples and thus can lead to a higher accuracy, at the expense of the training time. The basic Skip-Gram formulation defines $p(w_{t+j}|w_t)$ using the *softmax* function:

$$p(w_O|w_I) = \frac{\exp({v'_{wO}}^{\top} v_{wI})}{\sum_{w}^{W=1}\exp({v'_{w}}^{\top} v_{wI})} \tag{2}$$

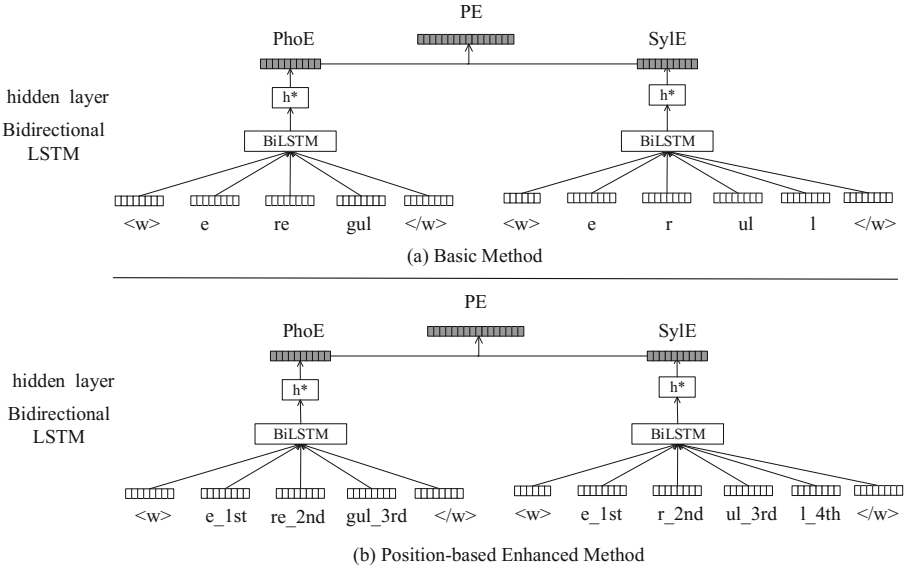

Fig. 2. (a) Basic method, (b) Position-based enhanced method for PE. Take a Mongolian word "eregul" for example, it contains three syllables: 'e', 're', 'gul' and four phonemes: 'e', 'r', 'ul', 'l'. (PhoE: phoneme embeddings; SylE: syllable embeddings)

where v_w and v'_w are the "input" and "output" vector representations of w, and W is the number of words in the vocabulary. This formulation is impractical because the cost of computing $\bigtriangledown \log p(w_O|w_I)$ is proportional to W, which is often large (10^5–10^7 terms).

2.1.2 Phonological Embeddings

PE are abstracted from these two vectors: PhoE and SylE. Figure 2 shows the process of obtaining the PE, through BiLSTM embedding network architecture, by using two methods in detail. We investigate two embedding methods for PE as shown in Fig. 2(a) and (b). We denote the two methods as "*Basic Method*" and "*Position-based Enhanced Method*".

Basic Method. In Fig. 2(a), each words in a Mongolian sentences is broken down into individual smaller phonological unit: phoneme and syllable, these are then mapped to a sequence of vectors $(v_1, \ldots, v_t)$, which are passed through a BiLSTM:

$$\overrightarrow{h_i^*} = LSTM(v_i, \overrightarrow{h_{i-1}^*}) \tag{3}$$

$$\overleftarrow{h_i^*} = LSTM(v_i, \overleftarrow{h_{i-1}^*}) \tag{4}$$

We then use the last hidden vectors from each of the LSTM components, concatenate them together, and pass the result through a separate non-linear layer.

$$h^* = [\overrightarrow{h_R^*}; \overleftarrow{h_l^*}] \qquad PhoE(SylE) = \tanh(W_m h^*) \tag{5}$$

where W_m is a weight matrix mapping the concatenated hidden vectors from both LSTMs into output representation *PhoE* or *SylE*. We then concatenate the PhoE and SylE to obtain PE.

Position-Based Enhanced Method. In the case of Mongolian phonemes and syllable, the same token holds different pronunciation information in different contexts. General *PhoE* and *SylE* in Fig. 2(a) cannot distinguish between uses of the same phoneme and syllable in different contexts. Motivated by [32], we add a positional tag for each phoneme and syllable in each word. Instead of keep three embeddings for each token corresponding to its three types of positions in a word, i.e., Begin, Middle and End. We add the novel positional tag, e.g. the **first/second/third/etc.** as shown in Fig. 2(b), according to the count of phonemes and syllables within a word, and then allow the model to learn more exquisite phonologically information for each specific word.

As demonstrated in Eqs. (3) and (4), we take a word (w_i) and its phonemes and syllables, $(p_1, \ldots, p_t)$ or $(s_1, \ldots, s_t)$, for example. We will take different embeddings of a phoneme and syllable according to its position within (w_i). That is, when building the embeddings (v_i), we will take the embeddings (v_first_i) for the beginning phoneme p_1 or syllable s_1 of the word w_i, take the embeddings (v_second_i) for the second phoneme p_2 or syllable s_2, and take the embeddings (v_third_i) for the third phoneme p_3 or syllable s_3. The rest is similar. Hence, Eqs. (3) and (4) can be rewritten as if we rename the new embeddings v_i as v_pos_i:

$$\overrightarrow{h_i^*} = LSTM(v\_pos_i, \overrightarrow{h_{i-1}^*}) \tag{6}$$

$$\overleftarrow{h_i^*} = LSTM(v\_pos_i, \overleftarrow{h_{i-1}^*}) \tag{7}$$

In the position-based enhanced phoneme and syllable embedding method, we learn new phoneme and syllable embeddings from positional tagged sequence of phonemes and syllables for each word. Various embeddings of each phoneme and syllable are differentiated by the position in the word, and the embedding assignment for a specific phoneme and syllable in a word can be automatically determined by the position. The position-based enhanced PE is obtained by concatenating the new PhoE and SylE.

2.1.3 Character Embeddings

CE are obtained using the same BiLSTM as described in Sect. 2.1.2. The BiLSTM takes as input a sequence of character of each Mongolian word and the last hidden states are used to create CE for the input word.

2.2 Attention Mechanism

We now have three alternative feature representation for each word - WE_t is an embedding learned on the word level as described in Sect. 2.1.1, PE_t is a representation dynamically built from the individual unit in the *t-th* word of the input Mongolian text, and CE_t is a representation learned from character unit as described in Sect. 2.1.3. Motivated by [33], instead of concatenating PE with the WE, the two vectors are added together using a weight sum, where the weights are predicted by a two-layer network:

$$w = \sigma(M_z^{(3)} \tanh(M_z^{(1)} \cdot WE + M_z^{(2)} \cdot PE)) \tag{8}$$

$$WE^* = w \cdot WE + (1 - w) \cdot PE \tag{9}$$

where $M_z^{(1)}$, $M_z^{(2)}$ and $M_z^{(3)}$ are weight metrics for calculating w, and $\sigma()$ is the logistic function with values in the range [0, 1]. The vector w has the same dimensions as WE or PE, acting as the weight between the two vectors. It allows the model to dynamically decide how much information to use from the phonologically component or the word embedding.

We concatenate the WE_t and PE_t into a complex vector (WE^*) using attention module, and then append the CE_t to WE^* to generate the joint embeddings (JE) as a new word-level representation for the phrase break prediction model: $JE = [WE^*; CE]$.

2.3 BiLSTM Phrase Break Model

As shown in Fig. 1. JE which concatenate word, phonological and character embeddings are given as input to two LSTM components moving in opposite directions through the text, creating context-specific representations. The respective forward- and backward-conditioned representations are concatenated for each word position, resulting in representations that are conditioned on the whole sequence:

$$\overrightarrow{h_t} = LSTM(JE_t, \overrightarrow{h_{t-1}}) \tag{10}$$

$$\overleftarrow{h_t} = LSTM(JE_t, \overleftarrow{h_{t-1}}) \tag{11}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{12}$$

We include an extra narrow hidden layer on top of the LSTM, which allows the model to detect higher-level feature combinations, while constraining it to be small forces it to focus on more generalisable patterns:

$$d_t = \tanh(W_d h_t) \tag{13}$$

where W_d is a weight matrix between the layers, and the size of d_t is intentionally kept small.

Finally, to produce phrase break label predictions, we use a softmax layer. An alternative approach for the output layer is using a CRF layer as in [34],

but we find that softmax output layer yields better results in our experiments. The softmax calculates a normalised probability distribution over all the possible labels of each word:

$$P(y_t = k|d_t) = \frac{e^{W_{o,k}d_t}}{\sum_{\widetilde{k}\epsilon K} e^{W_{o,\widetilde{k}}d_t}} \tag{14}$$

where $P(y_t = k|d_t)$ is the probability of the label of the t-th word (y_t) being k, K is the set of all possible labels, and $W_{o,k}$ is the k-th row of output weight matrix W_o. To optimise this model, we minimise categorical cross entropy, which is equivalent to minimising the negative log-probability of the correct labels:

$$E = -\sum_{T}^{t=1} \log(P(y_t|d_t)) \tag{15}$$

This approach assumes that the word-level, phonological-level components learn somewhat disjoint information, and it is beneficial to give word embedding only as input to the Mongolian phrase break prediction system. It allows the model to take advantages of phonological information from the phoneme and syllable in Mongolian.

3 Experiments and Analysis

3.1 Datasets

In the experiments, each word in a sentence was assigned to one of the following two PB labels: "B" and "NB" means *"break after a word"* and *"non-break"* respectively. For evaluating the effectiveness of the proposed approach, we rely on a corpus corresponding to the TTS database recorded by a professional native Mongolian female speaker. The corpus contains 59k sentences, more than 409k words, 1065k syllables and 1885k phonemes. The whole corpus is partitioned into training and test set for all experiments according to 4:1.

The word embedding train data were crawled from mainstream websites in Mongolian. After cleaning web page tags and filtering longer sentences, its token size and vocabulary are about 200 million and 3 million respectively.

3.2 Setup

For data preprocessing, all digits were replaced with the character "0". Any words that occurred only once in the training data were replaced by the generic OOV token for word embeddings, but were still used in the phonological embedding components. All tokens are initialized with 200 dimensional pre-trained vectors as illustrated in Sect. 2.1.1 and updated during training. The embeddings for phoneme and syllable parts were set to length 100 respectively and initialised randomly.

The LSTM layer size was set to 200 in each direction for different level components. The hidden layer d has size 50. Parameters were optimised using *AdaDelta*

with default learning rate 1.0 and sentences were grouped into batch of size 64. *Softmax* was used as the output layer for all the experiments. Performance on the training set was measured at every epoch and training was stopped if performance had not improved for 7 epoches; the best-performing model on training stage was then used for evaluation on the test set. Results are reported on the test set in terms of the Precision (P), Recall (R) and F-score (F) which is defined as the harmonic mean of the P and R.

3.3 Main Results

With this experiment, all Mongolian phrase break prediction systems are built at different input features. Table 1 report results from ablation tests experiments, with or without *Attention mechanism* (Att) or *Position-based enhanced method* (Pos), on various input features in Mongolian.

Our proposed method performs better than other benchmarks with the optimal configuration. Firstly, word embeddings alone perform rather poorly due to the challenges of reliably estimating them for a large vocabulary given a small dataset. Attach the PhoE or SylE to the word embeddings provide a significant performance boost. This indicates that phonological information benefits the Mongolian phrase break prediction from phoneme or syllable components. Using 'WE+PE' yields a further improvement in F score. Secondly, usage of attention mechanism seems to improve performance in all system in Table 1. This indicates that the attention mechanism is able to focus on the most effective information (word or phonological) adaptive to each token to maximize the information gain. Thirdly, results show the effectiveness of the position-based enhanced method for all model as well. In most cases, position-based phoneme and syllable embeddings can effectively distinguish different meanings of a word, which indicates, (1) modeling multiple senses of phonemes and syllables are important for phonological embeddings; (2) position information is adequate in addressing ambiguity. Among these, the proposed joint embedding ('WE+PE+CE') reaches the best performance for the usage of phonological information as well as the capability to fix the OOV problem.

3.4 Comparison of Phonological Embeddings Dimensions

In this experiment, we study the effect of varying the phonological embedding dimension on the performance of the model for Mongolian phrase break prediction. We use the best system ('WE+PE+CE' system with attention mechanism and position-based enhanced method) as described in Sect. 3.3 with all the parameters and hyperparameters unchanged, except for the dimension of PE (WE has the same dimensions as PE). We vary the PE dimension and compute the performance, in terms of F-Score. Figure 3 shows the results of this experiment.

Here we can see an indistinctive change in the performance on phrase break prediction when the PE dimension is varied. As it show in Fig. 3, the 200 dimension PE reach the best, while too much dimension will include other boring

Table 1. Ablation tests on different input features for Mongolian phrase break. (Att: Attention Mechanism; Pos: Position-based Enhanced Method)

Input features	Model	Att	Pos	P	R	F
WE	DNN [17]	No	No	86.92	82.20	82.95
	LSTM [17]	No	No	87.12	85.41	86.26
	BiLSTM	No	No	88.73	90.24	88.58
WE+PhoE	BiLSTM	No	No	91.13	90.58	89.94
		No	Yes	90.53	91.01	90.12
		Yes	No	90.24	90.96	90.04
		Yes	Yes	90.12	91.33	90.20
WE+SylE	BiLSTM	No	No	91.15	90.15	89.79
		No	Yes	91.09	90.17	90.03
		Yes	No	91.07	90.70	90.06
		Yes	Yes	90.18	91.27	90.13
WE+PE	BiLSTM	No	No	91.47	90.46	90.04
		No	Yes	90.32	91.03	90.23
		Yes	No	90.15	91.07	90.19
		Yes	Yes	90.83	91.43	90.40
WE+PE+CE	BiLSTM	No	No	91.04	91.21	90.27
		No	Yes	90.86	91.49	90.41
		Yes	No	90.85	91.38	90.39
		Yes	Yes	90.98	91.73	**90.82**

information that classifier cannot utilize, too small dimension can not learn the enough information. This is a somewhat surprising and counterintuitive result, as one would expect at least 1–2% increase in the performance corresponding to the increase in the PE dimensions.

3.5 Comparison of Position-Based Enhanced Method

In this experiment, we compare the performance of the position-based enhanced method by using two tagging schemes. A phoneme or syllable usually plays different roles when it is in different positions within a word. Here we utilize multiple-prototype phonological embeddings to address this issue. The idea is that, we keep multiple vectors for one phoneme and syllable, each corresponding to one of the meanings. We compare two tagging schemes in position-based enhanced method for multiple-prototype PE: (1) Tag location index for each token, e.g. the first/second/third/etc., according to the count of phonemes and syllables within a word; and (2) Keep three embeddings for each tokens corresponding to its three types of positions in a word, i.e., Begin, Middle and End. The first scheme, used in previous experiments, is named after 'Pos'. We denote

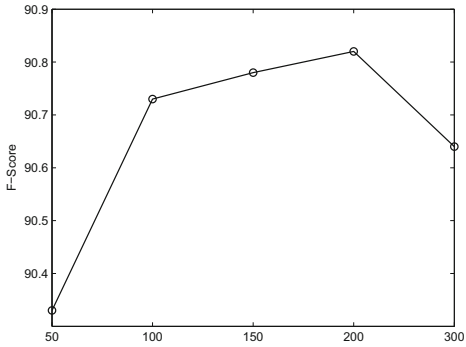

Fig. 3. Effect of varying the PE dimension on the performance (in terms of the F Score) of the 'WE+PE+CE' system on the Mongolian phrase break prediction.

Table 2. Effect of different position-based enhanced method on the performance of the 'WE+PE+CE' system for the Mongolian phrase break prediction.

Input features	Model	Att	Pos	Pos(BME)	P	R	F
WE+PE+CE	BiLSTM	No	No	Yes	90.49	90.92	90.29
		No	Yes	No	90.86	91.49	90.41
		Yes	No	Yes	90.74	91.51	90.75
		Yes	Yes	No	90.98	91.73	90.82

the second scheme as 'Pos(BME)'. Compared to the Basic Method (Sect. 2.1.2), with the two methods the count of (phoneme, syllable) can be increased from (61, 200) to (158, 378) and (546, 434) respectively.

We also use the 'WE+PE+CE' system as described in Sect. 3.3 with all the parameters and hyperparameters unchanged. Table 2 presents the results of this experiment. As examination of the results in Table 2 shows that 'Pos' method significantly outperform the 'Pos(BME)' method with or without attention mechanism. This proves that capturing more exquisite phonologically information using the scheme according to the count of phonemes and syllables within a word improves the performance, as compared to the 'Pos(BME)' method. Armed with 'Pos' method, the model can learn more rich phonologically information.

3.6 Comparison of Output Layer

In this experiment, we study the effect of different output layer on the performance for phrase break prediction. For sequence labeling task, it is effective to consider the correlation between taggers and jointly decode the best output for a given tokens. Following [34], we can also use a CRF as the output layer, which conditions each prediction on the previously predicted label. We denote the system using softmax or CRF output layer as 'BiLSTM-softmax' and 'BiLSTM-CRF'.

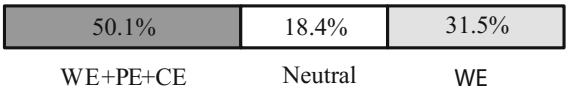

Fig. 4. The percentage preference of Subjective Evaluations.

Table 3 summarizes the experimental results for best system ('WE+PE+CE') by using BiLSTM-softmax or BiLSTM-CRF model. Surprisingly, we observe that BiLSTM-softmax model outperforms BiLSTM-CRF model in all metrics. This results conflicts with the empirical fact that CRF output layer is likely to work better than softmax output layer in other sequential labeling tasks, e.g., Part-of-Speech (POS) tagging [35] and Name Entity Recognition (NER) [36]. We believe that the abnormal phenomenon may owing to the nature of the specific tasks. For NER, which gives every word a entity label (e.g., time, location, organization, person, money, ...), the distribution of different labels in the corpus is approximately equal due to the grammar rules. It is important to consider the previous and future entity labels when decoding the current label. On the contrary, the distribution of output labels in phrase break prediction task is highly unequal, the proportion of 'NB' and 'B' is 85% versus 15% in our corpus. Thus the CRF layer may learn more knowledge about transiting from 'NB' to 'NB', and it will be more likely to predict the output label to be 'NB'. Modeling the output label dependencies, by the addition of a CRF layer on top of an BiLSTM, does not improve performance for Mongolian phrase break prediction model.

Table 3. Effect of BiLSTM model with different output layer on the performance of the WE+PE+CE system on the Mongolian phrase break prediction.

Input features	Model	P	R	F
WE+PE+CE	BiLSTM-softmax	90.98	91.73	90.82
	BiLSTM-CRF	89.67	89.08	89.69

3.7 Listening Tests

To compare the performance of the 'WE+PE+CE' system vs 'WE' one, a subjective preference listening test was conducted. A set of 20 sentence pairs of each session was randomly selected from the 100 pairs with different phrase break prediction results and speech was generated through a DNN-based Mongolian TTS system [37]. A group of 10 subjects were asked to choose which one was better in terms of the naturalness of synthesis speech, they could choose "Neutral" if they did not have any preference. The percentage preference is shown in Fig. 4. We can clearly see that the 'WE+PE+CE' system can achieve better naturalness of synthesized speech as compared with 'WE' system.

4 Conclusions

In this paper, we introduce a BiLSTM model with attention mechanism which uses the word embeddings, phonological representations and character embeddings to achieve state of the art performance. Our experiments show that including a phonological representation component in the BiLSTM model provides substantial performance improvements on all the benchmarks for Mongolian phrase break prediction. In addition, the attention mechanism and position-based enhanced method for phonological representations achieve the best results on all evaluations. Moreover, all of this is achieved without any extra feature engineering specific to the task or language. Our method can also be applied to various languages.

References

1. Chen, Z., Hu, G., Jiang, W.: Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction. In: 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 1421–1424 (2010)
2. Chu, M., Qian, Y.: Locating boundaries for prosodic constituents in unrestricted mandarin texts. Comput. Linguist. Chin. Lang. Process. **6**, 61–82 (2001)
3. Nie, X., Wang, Z.: Automatic phrase break prediction in Chinese sentences. J. Chin. Inf. Process. **17**(4), 39–44 (2003)
4. Li, J.F., Hu, G.P., Wang, R.: Chinese prosody phrase break prediction based on maximum entropy model. In: 8th Proceedings of INTERSPEECH, Jeju Island, Korea, pp. 729–732 (2004)
5. Qian, Y., Wu, Z., Ma, X., Soong, F.: Automatic prosody prediction and detection with conditional random field (CRF) models. In: 7th Proceedings of ISCSLP, Tainan, Taiwan, pp. 135–138 (2010)
6. Rosenberg, A., Fernandez, R., Ramabhadran, B.: Phrase boundary assignment from text in multiple domains. In: 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, pp. 2558–2561 (2012)
7. Vadapalli, A., Bhaskararao, P., Prahallad, K.: Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for Indian languages. In: 8th ISCA Tutorial and Research Workshop on Speech Synthesis (2013)
8. Ananthakrishnan, S., Narayanan, S.: An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: 30th International Conference on Acoustics. Speech, and Signal Processing, pp. 269–272. IEEE Press, Philadelphia (2005)
9. Hasegawa-Johnson, M., et al.: Simultaneous recognition of words and prosody in the Boston University radio speech corpus. Speech Commun. **46**, 418–439 (2005)
10. Sridhar, V.K.R., Bangalore, S., Narayanan, S.S.: Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Trans. Audio Speech Lang. Process. **16**, 797–811 (2008)
11. Busser, B., Daelemans, W., van den Bosch, A.: Predicting phrase breaks with memory-based learning. In: 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire Scotland (2001)

12. Fernandez, R., Ramabhadran, B.: Driscriminative training and unsupervised adaptation for labeling prosodic events with limited training data. In: 11th Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, pp. 1429–1432 (2010)
13. Rosenberg, A., Fernandez, R., Ramabhadran, B.: Modeling phrasing and prominence using deep recurrent learning. In: 16th Conference of the International Speech Communication Association, Dresden, Germany, pp. 3066–3070 (2015)
14. Vadapalli, A., Prahallad, K.: Learning continuous-valued word representations for phrase break prediction. In: 15th Conference of the International Speech Communication Association, Singapore, pp. 41–45 (2014)
15. Watts, O., et al.: Neural net word representations for phrase-break prediction without a part of speech tagger. In: 34th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, pp. 2599–2603 (2014)
16. Watts, O., Yamagishi, J., King, S.: Unsupervised continuous-valued word features for Phrase-break prediction without a part-of-speech tagger. In: 12th Conference of the International Speech Communication Association, Florence, Italy (2011)
17. Vadapalli, A., Gangashetty, S.V.: An investigation of recurrent neural network architectures using word embeddings for phrase break prediction. In: 17th Conference of the International Speech Communication Association, San Francisco, CA, USA, pp. 2308–2312 (2016)
18. Rendel, A., Fernandez, R., Hoory, R., Ramabhadran, B.: Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end. In: 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, pp. 5655–5659 (2016)
19. Ding, C., Xie, L., Yan, J., Zhang, W., Liu, Y.: Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features. In: IEEE Automatic Speech Recognition and Understanding Workshop, Scottsdale, Arizona, USA, pp. 98–102 (2015)
20. Zheng, Y., Li, Y., Wen, Z., Ding, X., Tao, J.: Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach. In: 17th Conference of the International Speech Communication Association, San Francisco, CA, USA, pp. 3201–3205 (2016)
21. Klimkov, V., et al.: Phrase break prediction for long-form reading TTS: exploiting text structure information. In: 18th Conference of the International Speech Communication Association, Stockholm, Sweden, pp. 1064–1068 (2017)
22. Liu, R., Bao, F., Gao, G., Wang, W.: Mongolian prosodic phrase prediction using suffix segmentation. In: International Conference on Asian Language Processing, pp. 250–253. IEEE (2017)
23. Gertai, Q.: Mongolian Syntax, pp. 77–133. Mongolia People Publishing House, Hohhot (1991)
24. Temusurvn and Otegen: Mongolian Orthography Dictionary, pp. 77–133. Inner Mongolia People Publishing House, Hohhot (1999)
25. Bao, F., Gao, G., Yan, X., Wang, W.: Segmentation-based Mongolian LVCSR approach. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 8136–8139 (2013)
26. Ling, W., et al.: Finding function in form: compositional character models for open vocabulary word representation. Computer Science, pp. 1899–1907 (2015)
27. Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2016)

28. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (2002)
29. Mikolov, T., et al.: Efficient estimation of word representations in vector space. Computer Science (2013)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
31. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
32. Chen, X., Xu, L., Liu, Z., Sun, M., Luan, H.B.: Joint learning of character and word embeddings. In: International Conference on Artificial Intelligence, pp. 1236–1242, AAAI Press (2015)
33. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computer Science (2014)
34. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. Computer Science (2015)
35. Liu, R., Bao, F., Gao, G., Wang, Y., et al.: Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. In: 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), Taipei, Taiwan (2017)
36. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, pp. 260–270 (2016)
37. Liu, R., Bao, F., Gao, G., Wang, Y.: Mongolian text-to-speech system based on deep neural network. In: Tao, J., Zheng, T.F., Bao, C., Wang, D., Li, Y. (eds.) NCMMSC 2017. CCIS, vol. 807, pp. 99–108. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8111-8_10