Mandarin Prosody Prediction Based on Attention Mechanism and Multi-model Ensemble

Kun Xie(✉) and Wei Pan(✉)

Fujian Key Laboratory of Brain-Inspired Computing Technique and Applications, School of Information Science and Engineering, Xiamen University, Xiamen, China
xierhacker@stu.xmu.edu.cn, wpan@xmu.edu.cn

Abstract. Prosodic boundary prediction is very important and challenging in the speech synthesis task, the result of prosodic prediction directly determines the quality of speech synthesis. In this paper, we proposed a prosodic boundary prediction method based on "encoding-decoding" frame while using an effective position attention mechanism to further improve performance. Finally, we investigate the use of Random Forest and Gradient Boosting Decision Tree to explore the potential of combined multiple models. The experimental results show that compared with the current best method of prosodic structure (Bi-LSTM), the proposed method presented a good result with F1-Score in terms of prosodic words, prosodic phrases, intonation phrases; the subjective experiment also shows that the proposed method can improve the quality and naturalness of synthesized speech.

Keywords: Speech synthesis · Prosodic boundaries prediction
Attention mechanisms · Bi-LSTM · Model ensemble

1 Introduction

Speech synthesis is a very important field in the study of speech interaction, it's main purpose is to convert normal language text into speech. Generally, we use naturalness and intelligibility [1] to judge the quality of the Speech synthesis system. In recent years, the performance of speech synthesis is getting better and better, but the naturalness of synthesized speech still has a certain gap compared with the real speech. One of the most important reason is that the performance of the existing prosody prediction model is not good enough.

Prosody prediction can be treated as a sequence labeling or sequence to sequence problem, we want to learn a function $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ that maps an input sequence $\boldsymbol{x}$ to the corresponding label sequence $\boldsymbol{y}$. In fact, we add labels to indicate the boundary of each word so we can restore the prosodic structure through this boundary information. Therefore, the essence of the prosody prediction problem is a boundary judgment problem at different prosodic levels. In mandarin speech synthesis systems, a typical hierarchical prosodic structure can significantly improve the accuracy of prosody

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10954, pp. 491–502, 2018.
https://doi.org/10.1007/978-3-319-95930-6_45

prediction to distinguish different levels of pauses between words in speech. Normally, the prosodic boundaries are often classified into prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH) [2].

In this work, we investigate how prosody prediction can benefit from the strong modeling capacity of sequence to sequence models. we also investigate the use of multi-model ensemble to explore the performance of combined multiple models at the decision level. The rest of this paper is organized as follows: Sect. 2 offers a brief overview of the existing research work that is related to this research; in Sects. 3, 4 the proposed approach is described in detail; experimental results are given in Sect. 5 to demonstrate the feasibility and performance of the proposed method; and, finally, a brief conclusion and future works are presented in Sect. 6.

2 Related Works

2.1 Traditional Methods

Traditional methods including rules-based methods and statistical machine learning based methods. Most of the early methods are rules-based methods, the idea of which is to use empirical rules to map the syntactic structure to the level of the prosodic structure [3]. With the development of statistical machine learning, many machine learning methods are gradually applied to the prediction of prosodic structures, such as decision trees [2] and conditional random fields (CRF) [4].

It is worth mentioning that the CRF model is one of the best models that we still using now, it's a very simple but surprisingly effective tagging model. therefore, in comparison with the traditional methods, we will focus on the CRF model.

The core idea of CRF is not complicated, for sequence labeling tasks, it is often beneficial to explicitly consider the correlations between adjacent labels [28]. Correlations between adjacent labels can be modeled as a transition matrix $\boldsymbol{T} \in \mathbb{R}^{C\times C}$. Given a sentence $\boldsymbol{S} = (c_1, c_2, \ldots, c_L)$, we have corresponding scores $\boldsymbol{E} \in \mathbb{R}^{L\times C}$. For a label sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_L)$, we define it's unnormalized score to be

$$s(\boldsymbol{S}, \boldsymbol{y}) = \sum_{i=1}^{L} E_{i,y_i} + \sum_{i=1}^{L-1} T_{y_i,y_{i+1}} \tag{1}$$

then we can takes the form of linear chain CRF [29], the probability of the label sequence is defined as

$$P(\boldsymbol{y}|\boldsymbol{S}) = \frac{e^{s(\boldsymbol{S},\boldsymbol{y})}}{\sum_{\boldsymbol{y}'\in\boldsymbol{Y}} e^{s(\boldsymbol{S},\boldsymbol{y}')}} \tag{2}$$

where $\boldsymbol{Y}$ is the set of all valid label sequences. Then the loss of the proposed model is defined as the negative log-likelihood of the ground-truth label sequence y^*,

$$L(\boldsymbol{S}, \boldsymbol{y}^*) = -\log P(\boldsymbol{y}^*|\boldsymbol{S}) \tag{3}$$

During training, the loss function is minimized by back propagation. During test, Viterbi algorithm is applied to quickly find the label sequence with maximum probability.

2.2 Deep Neural Networks Based Method

As a sequence labeling problem, prosody prediction can be expressed like other sequence labeling problems as Eq. (4),

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}) \tag{4}$$

means that we want to find the best label sequence $\boldsymbol{y}$ given an input sequence $\boldsymbol{x}$.

In practical terms, we will considering all available information from the input and the emitted output sequences, it's turned to find the best parameter set θ that maximizes the likelihood which can be described by the following expression

$$\arg\max_{\boldsymbol{\theta}} \prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{y}_1^{t-1}, \boldsymbol{x}; \boldsymbol{\theta}) \tag{5}$$

where y_1^{t-1} represents the predicted output sequence prior to time step t.

With the rapid development of deep neural networks in recent years, related technologies and models have been applied to many fields [5–7], those methods also been applied to sequence modeling problems [8, 9]. Vadapalli et al. [10] proposed a prosodic structure prediction model based on RNN and added word vectors as semantic features. Experimental results show that RNN-based methods can greatly improve the performance comparing by traditional machine learning methods (such as conditional random fields), and meanwhile, the vector feature of words can well adapt to the model of cyclic neural network. Ding et al. [11] from another perspective, the word (rather than the traditional grammatical word) as the unit of prediction of prosodic structure, use of vector as the RNN input to replace other traditional features directly. The advantage of this approach is it does not rely on the precision of other text analysis.

Recently, encoder-decoder neural network frames have been successfully applied in many sequence learning problems such as machine translation [8] and speech recognition [12]. The framework is firstly introduced in [8, 13] and the encoder and decoder are two separate RNNs. The main idea behind the encoder-decoder frame is to encode input sequence $\boldsymbol{x}$ into a dense vector $\boldsymbol{c}$. This vector encodes information of the whole source sequence and then use this vector to generate corresponding output sequence, which can be expressed as follows:

$$P(\boldsymbol{y}) = \prod_{t=1}^{T} P(\boldsymbol{y}_t|\boldsymbol{y}_1^{t-1}, \boldsymbol{c}) \tag{6}$$

The attention mechanism introduced in [14] enables the encoder-decoder architecture to learn to align and decode simultaneously.

3 Proposed Approach

3.1 Basic Encoder-Decoder Framework

The prosody prediction task is to generate the boundary labeling sequence $\boldsymbol{y}$ from the word sequence $\boldsymbol{x}$, Let $\boldsymbol{x}_i$ to represent a word and 0 or 1 represent the Prosodic boundary. Considering the ability to better model long-term dependencies, we use LSTM [15] as the basic recurrent network unit. Furthermore, bidirectional LSTM as the encoder module and unidirectional LSTM as the decoder module. Based on this, the main Framework is illustrated in Fig. 1.

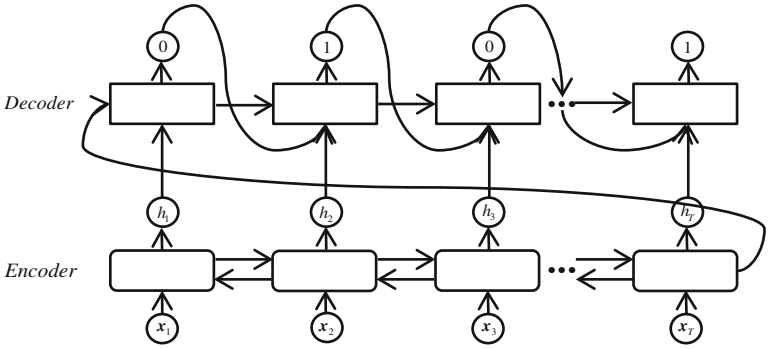

Fig. 1. Encoder-decoder model with aligned input

The encoder reads the word sequence forward and backward. The forward processes and the backward processes read the word sequence in its original order and reverse order, at the same time, generates forward hidden states and backward hidden states at each time step. The final encoder hidden state at each time step is a concatenation of the forward states and backward states. The last state of the encoder carries information of the entire input word sequence and we use it and the initial decoder hidden state. The decoder reads the hidden states sequence forward and generate boundary labeling sequence. At each decoding step, the decoder state is calculated as a function of the previous decoder state, the previous emitted label, and the aligned encoder hidden state.

3.2 Encoder-Decoder Framework with Attention

The attention mechanism is proposed mainly to solve the problem of losing hidden states information [8, 13]. Under the encoder-decoder frame, the hidden states carry information of the whole input sequence, but along the forward and backward propagation, information may gradually lose. Thus, instead of only utilizing the hidden state

at each step, the use of context vector $\boldsymbol{c}$ gives us any additional supporting information, especially those require longer term dependencies that is not being fully captured by the hidden state. Motivated by the use of attention mechanism in encoder decoder frames, we propose the attention-based model for prosody prediction which main framework is illustrated in Fig. 2.

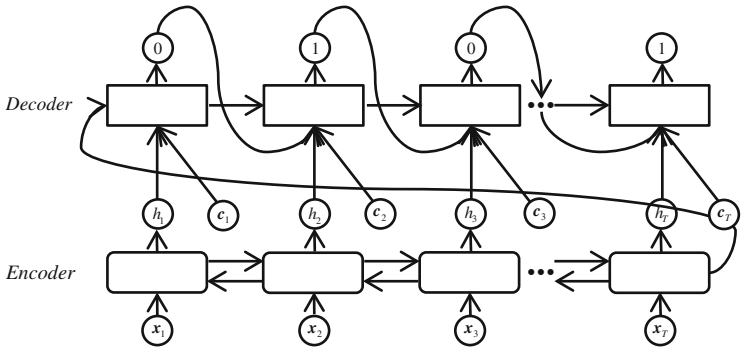

Fig. 2. Encoder-decoder model with aligned input and attention

In the proposed model, we still use the same basic encoder-decoder frame as above, encoder reads the word sequence forward and backward and generate hidden states. The only difference is that we add a context vector $\boldsymbol{c}_i$ at each time step. where the context vector $\boldsymbol{c}_i$ is calculated as a weighted average of the encoder hidden states $\boldsymbol{h} = (h_1, h_2, \ldots\ldots, h_T)$. the context vector can be regarded as information that when model focusing on a few hidden states. we reuse the pre-computed hidden states $\boldsymbol{h}$ of the encoder to produce intent class distribution.

There are many studies on the mechanism of attention [16, 17] and which have achieved very good results. We adopt positional attention [18] to calculate attention weights which is suitable for prosody prediction task.

We can describe that progress as follows:

$$e(\boldsymbol{h}_j, \boldsymbol{p}_j) = \boldsymbol{V}^T \tanh(\boldsymbol{W}_H \boldsymbol{h}_j + \boldsymbol{W}_P \boldsymbol{p}_j + \boldsymbol{b}) \tag{7}$$

And,

$$\alpha_j = \frac{\exp(e(\boldsymbol{h}_j, \boldsymbol{p}_j))}{\sum_{k=1}^{K} \exp(e(\boldsymbol{h}_k, \boldsymbol{p}_k))} \tag{8}$$

Finally, the context vector can be expressed by Eq. (9),

$$\boldsymbol{c}_i = \sum_{j=1}^{T} \alpha_j \boldsymbol{h}_j \tag{9}$$

4 Multi-model Ensemble

Ensemble learning achieve a better generalization performance than a single learner by constructing and combining multiple learners. According to the generation of individual learners, the current ensemble learning methods can be broadly divided into two categories, the first is Boosting, it is a strong dependency between the individual learners, serialization methods must be generated in series; the second is Bagging, there is no strong dependencies between individual learner and can be generated in parallel. We use Random Forest(RF) and Gradient Boost Decision Tree (GBDT), the two most representative Boosting and Bagging algorithms as the main method of model ensemble.

4.1 Random Forest

Random Forest, RF [19] is an extended variant of Bagging. The training algorithm for random forests applies the general technique of bagging to tree learners. Given a training set $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with responses $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples. After training, predictions for unseen samples $\boldsymbol{x}'$ can be made by averaging the predictions from all the individual regression trees on $\boldsymbol{x}'$:

$$\hat{f} = \frac{1}{T}\sum_{t=1}^{T} f_t(\boldsymbol{x}') \tag{10}$$

or by taking the majority vote in the case of classification trees. Random forest is simple and easy to implement, and leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees.

4.2 Gradient Boosting Decision Tree

Gradient Boosting Decision Tree, GBDT [20, 21] is a typical method belonging to Boosting. In the GBDT iteration, suppose the strong learner we obtained in the previous iteration is $f_{t-1}(\boldsymbol{x})$ and the loss function is $L(y, f_{t-1}(\boldsymbol{x}))$. The goal of our current iteration is to find a CART [22] $T_t(\boldsymbol{x})$ minimizes the loss,

$$L(y, f_t(\boldsymbol{x})) = L(y, f_{t-1}(\boldsymbol{x})) + T_t(\boldsymbol{x}) \tag{11}$$

A decision tree partitions the space of all joint predictor variant into disjoint regions R_j, $j = 1, 2, \ldots, J$. as represented by the terminal nodes of the tree. A constant γ_j is assigned to each such region. Then GBDT is a sum of such trees, where M is the numbers of trees in GBDT,

$$f_M(\boldsymbol{x}) = \sum_{m=1}^{M} T_m(\boldsymbol{x}) \tag{12}$$

GBDT is complicated than RF and can't be generated in parallel, but it may leads to a better performance in some tasks.

4.3 Structure

We adopt 4 classifier: CRF, BLSTM, Basic encoder-decoder frame(aligned model), Attention model as the basic classifier. The framework of multi-model ensemble is shown in Fig. 3.

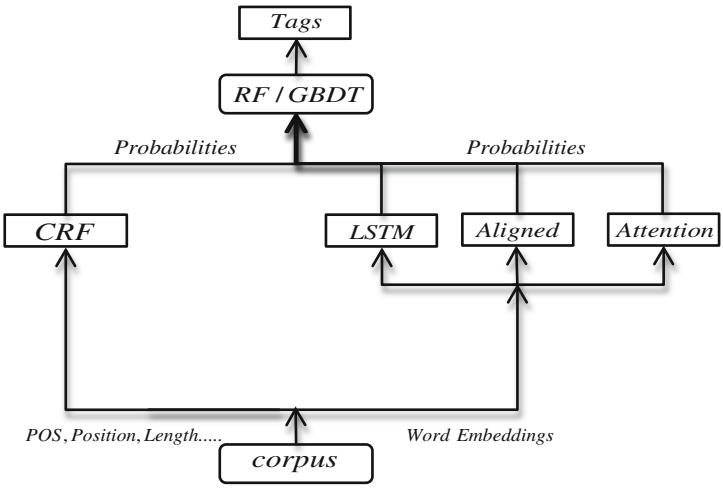

Fig. 3. Flowchart of model ensemble

Firstly, the CRF uses linguistic class features (POS, Position information etc.) and the LSTM, Aligned mode, Attention model use embedding features, then the probability of Breaks (PW, PPH and IPH) can be obtained by these four single classifiers; Next, the output probabilities, together with the important features are consisting of the inputs for model fusion module. During model fusion, two different methods, RF and GBDT are trained and employed to make the final prediction.

5 Experiments

5.1 Settings

Considering that the word is often used as the ideographic unit in Mandarin, which carries more semantic and boundary information than isolated character, we decide use word as basic input units in our experiments.

Datasets. We evaluate our model on a speech synthesis corpus which contains 66150 sentences. The whole corpus is partitioned into training, validation and test set for all experiments according to 7:2:1. Word segmentation and POS tagging are also annotated by expert annotator.

Word Embeddings. Word embeddings [23, 24] lead to a significant improvements over the linguistic features which don't take into account the distributional behavior of words [25]. And this issue has been addressed by word embedding which encodes a word as a low-dimensional vector. We used 8G Mandarin text corpus which word vocabulary size is 393255, The word embedding dimension is set as 128, the context window size is set as 5 during training, and we also use both hierarchical softmax and 5-word negative sampling.

CRF. Same as traditional labeling tasks, we train CRF basically on linguistic features like POS, word position, the length of word, etc.

Bi-LSTM. Standard method used for prosodic boundaries prediction based on Bi-LSTM. We use 2-layer Bi-LSTM as the main time series modeling part. each layer with 256x2 LSTM hidden units, and then followed a fully connected layer, finally, a binary output softmax layers is used to output the probabilities of each boundary.

Aligned. Basic encoder-decoder frame with aligned input show in Fig. 2. Encoder is a 1-layer Bi-LSTM with 256x2 hidden units. Decoder is a 1-layer unidirectional LSTM with 256 hidden units. the output of encoder as the input of decoder, and the final state of encoder as the initial state of decoder.

Attention. Encoder-decoder frame with attention show in Fig. 3. The main structure is the same as LSTM_Aligned, the only difference is that it adds attention mechanism.

RF. Random forest for prosodic boundaries prediction based on the output probability of the four single classifiers (CRF, BLSTM, LSTM_Aligned and LSTM_Attention).

GBDT. GBDT for prosodic boundaries prediction based on the output probability of the four single classifiers (CRF, BLSTM, LSTM_Aligned and LSTM_Attention).

All networks mentioned above are trained with Adam [26], the initial learning rate been set to 0.003 and reset the learning rate by exponential decay (decay rate set to 0.02) for every epoch. The batch size is 20.

5.2 Main Result

Table 1 gives the performances of all six methods described in 5.1, We report our results in terms of F1-Score, which is defined as the harmonic mean of precision and recall. We analyze the results below.

First, we compare standard statistical model CRF with the standard neural model LSTM. In Table 1, we can see clearly that the Bi-LSTM model perform well in all the hierarchies. Many successful research and experiences also shown that LSTM-based model performs better than statistical methods for complex annotation tasks.

Then, we compare Bi-LSTM, Aligned model and Attention model. These methods are all based on LSTM, the differences is the structure and whether the attention

Table 1. Performance of F1-Score

	CRF	Bi-LSTM	Aligned	Attention	RF	GBDT
PW	95.02	95.61	96.52	96.76	97.13	**97.22**
PPH	82.04	82.25	83.82	83.86	84.84	**84.91**
IPH	78.80	80.42	83.49	83.62	83.86	**83.89**

mechanism been applied. We can infer from the Table 1 that the aligned encoder-decoder structure perform better than basic Bi-LSTM, we suspect that compared to the use of LSTM to do classification, the prosody prediction task is more biased towards the seq2seq architecture. Not to our surprise, the model with attention perform best in all of them, and, the higher the prosodic level (such as IPH), the more obvious the effect of improvement. The probable reason for this is that for IPH, a longer span of contextual information is needed to make the correct prediction. This is the same with IPH labeling, the same need to consider the entire sentence information to give the final annotation results. With respect to aligned based encoder-decoder, attention-based encoder-decoder is precisely this predictive-enabling contextual information.

Finally, we can see from Table 1 that the integrated model (RF and GBDT) achieves superior performance than all the single models and GBDT shows an absolute increase of around 5% than CRF for IPH prediction. It can be explained by ensemble learning can learn the advantages of all the individual models, prevent over-fitting effectively and improve the generalization performance. In the comparison between GBDT and RF, GBDT performed slightly better, in fact the performance of the two are very close, all showed the ability of ensemble learning should have.

Considering that the linguistic class features can be also applied into the deep learning models, if we combine the word embeddings and all the linguistic feature together we can get a new word embeddings. Specifically, we concatenate the word POS tag, position, length, cumulative length and original word embeddings by last dimension and use this new word embeddings to train and test all the models, the result shown in Table 2.

Table 2. Performance of F1-Score on muilti-features

	CRF	Bi-LSTM	Aligned	Attention	RF	GBDT
PW	95.02	95.75	97.03	97.11	97.25	**97.26**
PPH	82.04	82.31	84.21	84.31	85.01	**85.06**
IPH	78.80	80.46	83.63	83.70	83.92	**83.96**

By comparing Tables 1 and 2, we can see a slight improvement in the performance of each model, and the convergence of multiple features is helpful for the final performance improvement. To evaluate the effects of four single classifiers for the best ensemble method GBDT. We further calculate the contribution of each feature by Gini importance [27], which is used as a general indicator of feature importance. Take IPH

prediction for example, the degree of the top five features contributions for IPH prediction on word-based GBDT are listed in Table 3.

This means the output from the LSTM based method are playing the dominant role rather than CRF. The previous results in Table 1, we can see that the Attention-based approach can lead to more accurate results, and it does contribute the most to all methods in this ensemble method.

Table 3. Model contribution in GBDT

Methods	Contributions
CRF	12.96%
Bi-LSTM	22.43%
Aligned	31.77%
Attention	**32.84%**

6 Conclusions

In this paper, we explored strategies in utilizing explicit alignment information in the attention-based encoder-decoder neural network models for Mandarin prosodic boundaries prediction, and we also use ensemble learning to further enhance the generalization performance of this model. Our results show that attention-based model is more suitable than basic LSTM approach, compared to a single model, the use of multi-model ensemble can bring very large performance improvements. Meanwhile, the model fusion results indicate that the dependency of results on BLSTM is greater than CRF, and the features generated from feature ranking module can further boost the performance of prosodic boundaries prediction. In the future, We have two paths to go, the first is try to deepen the encoders and decoders to see if it can learn more abstract concepts. The second is to change the type of encoder and decoder, for example, we can use CNN to be the encoder or the decoder.

References

1. Taylor, P.: Text-to-Speech Synthesis. Cambridge University Press, Cambridge (2009)
2. Chu, M., Qian, Y.: Locating boundaries for prosodic constituents in unrestricted Mandarin texts. Int. J. Comput. Linguist. Chin. Lang. Process. **6**(1), 61–82 (2001). Special Issue on Natural Language Processing Researches in MSRA
3. Truckenbrodt, H.: Phonological phrases–their relation to syntax, focus, and prominance. Massachusetts Institute of Technology (1995)
4. Levow, G.A.: Automatic prosodic labeling with conditional random fields and rich acoustic features. In: Proceedings of the Third International Joint Conference on Natural Language Processing, vol. I (2008)
5. Geng, Y., Liang, R.Z., Li, W., Wang, J., Liang, G., Xu, C., Wang, J.Y.: Learning convolutional neural network to maximize pos@top performance measure. In: ESANN 2017 - Proceedings, pp. 589–594 (2016)

6. Geng, Y., Zhang, G., Li, W., Gu, Y., Liang, R.Z., Liang, G., Wang, J., Wu, Y., Patil, N., Wang, J.Y.: A novel image tag completion method based on convolutional neural transformation. In: International Conference on Artificial Neural Networks, pp. 539–546 (2017)
7. Zhang, G., Liang, G., Li, W., Fang, J., Wang, J., Geng, Y., Wang, J.Y.: Learning convolutional ranking-score function by query preference regularization. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 1–8 (2017)
8. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
9. Mikolov, T., Kombrink, S., Burget, L., et al.: Extensions of recurrent neural network language model. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5528–5531. IEEE (2011)
10. Vadapalli, A., Prahallad, K.: Learning continuous-valued word representations for phrase break prediction. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
11. Ding, C., Xie, L., Yan, J., et al.: Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 98–102. IEEE (2015)
12. Chan, W., Jaitly, N., Le, Q., et al.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE (2016)
13. Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
14. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
16. Li, H., Min, M.R., Ge, Y., et al.: A context-aware attention network for interactive question answering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 927–935. ACM (2017)
17. Wang, X., Yu, L., Ren, K., et al.: Dynamic attention deep model for article recommendation by learning human editors' demonstration. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2051–2059. ACM (2017)
18. Chen, Q., Hu, Q., Huang, J.X., et al.: Enhancing recurrent neural networks with positional attention for question answering. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 993–996. ACM (2017)
19. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
20. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
21. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
22. Breiman, L., Friedman, J., Stone, C.J., et al.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
23. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)

24. 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)
25. Watts, O., Yamagishi, J., King, S.: Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980 (2014)
27. Menze, B.H., Kelm, B.M., Masuch, R., et al.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinf. **10**(1), 213 (2009)
28. Collobert, R., Weston, J., Bottou, L., et al.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
29. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)