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

Pre-trained Text Representations for Improving Front-End Text Processing in Mandarin Text-to-Speech Synthesis

Bing Yang, Jiaqi Zhong, Shan Liu

Cloud and Smart Industries Group, Tencent Technology Co., Ltd. China
{bjarneyang, jackiezhong, shiningliu}@tencent.com

Abstract

In this paper, we propose a novel method to improve the performance and robustness of the front-end text processing modules of Mandarin text-to-speech (TTS) synthesis. We use pre-trained text encoding models, such as the encoder of a transformer based NMT model and BERT, to extract the latent semantic representations of words or characters and use them as input features for tasks in the front-end of TTS systems. Our experiments on the tasks of Mandarin polyphone disambiguation and prosodic structure prediction show that the proposed method can significantly improve the performances. Specifically, we get an absolute improvement of 0.013 and 0.027 in F1 score for prosodic word prediction and prosodic phrase prediction respectively, and an absolute improvement of 2.44% in polyphone disambiguation compared to previous methods.
Index Terms:Text to Speech Front-End, Polyphone Disambiguation, Prosodic Structure Prediction, Pre-trained Text Representation

1. Introduction

Recently, great progress has been made in the field of text-to-speech (TTS). The speech synthesized by recently proposed end-to-end acoustic models (e.g. Tacotron [1], transformer TTS [2], etc) and neural vocoders (e.g. WaveNet [3], WaveRNN [4], WaveGlow [5] etc) is almost undistinguishable from recorded human speech. However, the text processing module, a.k.a. the front-end, still plays an important role in the TTS system, especially in Mandarin TTS, since the limited training data in most of the TTS tasks is unable to cover the variety in input text to be synthesized.

The front-end of TTS aims to extract various linguistic and phonetic features from the raw text, in order to improve the naturalness and intelligibility of the synthesized speech. The front-end of a Mandarin TTS system contains a series natural language processing (NLP) modules, including text normalization (TN) [6], Chinese word segmentation (CWS) [7], part-of-speech (POS) tagging [8], polyphone disambiguation (PPD) [9] and prosodic structure (PS) prediction [10] etc.

Previous researches on the TTS front-end can be divided into two categories. One is the traditional statistical methods, such as maximum entropy (ME) [11, 12, 13], conditional random field (CRF) [14] and Classification And Regression Tree (CART) [15], using manually designed linguistic features, e.g. POS, word-terminal syllables, word context etc. The main problem of these methods is that they require the knowledge of linguistic experts to design task relevant features. The other category includes recurrent neural network (RNN) based methods [16, 17, 18]. An RNN with gated cells, e.g. long-short term memory (LSTM) [19] and gated recurrent unit (GRU) [20], is a sequential model which has been successfully applied to speech and NLP tasks. These RNN based models can be learned in an end-to-end manner with manually designed feature. The main drawback of RNN based method is that it usually requires large amount of training data to achieve good performance and generalization. However, the amount of training data in most of the TTS front-end tasks is usually very limited. Therefore, RNNs usually work with pre-trained word-level or character-level embedding representations (e.g. word2vec [21]) to improve the generalization of the learned models. Another existing problem is that long-term context dependency [22] is usually required in the TTS front-end tasks. And this dependency can not be well captured by using the previous methods.

To address these issues of the previous methods, we propose to use stronger text representation extractors to improve the performance of our front-end tasks. In this work, we use the bidirectional encoder representations from transformers, a.k.a. the BERT [23], a recently proposed NLP pre-training method, and the encoder of a transformer [24] based neural machine translation (NMT) system [25] to extract latent text representations with semantic meanings and use them as input features into task specific models. Both of these two feature extractors utilize a deep multi-head self-attention structure to capture the long-term context dependency in text sequences. In our experiments, we applied the extracted representation to the polyphone disambiguation task, as well as the prosodic word (PW) and prosodic phrase (PP) prediction tasks. We adopted a multi-layer feed-forward neural network to predict the correct pronunciation from the representation of a polyphone character. We also used an LSTM-based bidirectional RNN-CRF [26, 27, 28] to predict PP and PW tags using the representation sequence of a sentence as input. Our experiments show that both BERT and NMT encoder can help the model to achieve significant improvement compare with conventional methods. Moreover, benefiting from the powerful representation ability of the pre-trained models, we found that PP and PW prediction can be learned in a single network framework without losing any accuracy by using a multi-task training method.

This paper is organized as follows: Section 2 will briefly review the basic background of front-end of Mandarin TTS. The proposed method will be given in Section 3. Experimental details and results will be given in Section 4. Lastly in Section 5, some conclusions and potential future research will be given.

2. Mandarin TTS Front-End

In this section, we will briefly explain the main tasks of the front-end of Mandarin TTS systems and the conventional approaches to these tasks. The goal of the front-end of a TTS system is to extract phonetic and linguistic information require by back-end acoustic models from the input raw text. Figure 1 shows a typical pipeline of a Mandarin TTS text processing front-end. It usually contains three main components, including text normalization, prosodic structure prediction and grapheme-

http://dx.doi.org/10.21437/Interspeech.2019-1418

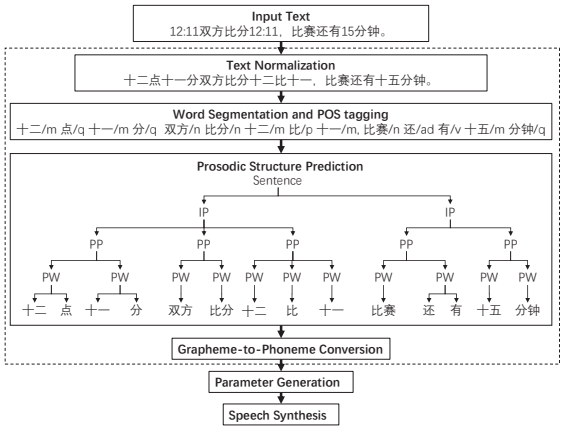

Figure 1: *A typical pipeline of text processing of a Mandarin text-to-speech.*

to-phoneme (G2P) conversion [29, 30].

2.1. Text normalization

The text normalization module aims to convert written raw text into its spoken form. E.g. TN should convert input text "10.5%" in "ten point five percent". Rule-based methods, which utilize regular expression matching, is usually adopted to perform this conversion. Recently, TN is also treated as a sequence-to-sequence problem handled by an attention-based RNN [18] .

2.2. Polyphone disambiguation

The G2P conversion for Mandarin characters mainly focuses on polyphone disambiguation. There are thousands of characters in Mandarin text, among which about one thousand characters are polyphone characters. This means that each of these characters has multiple pronunciations. Polyphone disambiguation is critical for Mandarin TTS since different pronunciation always results in different meaning of a sentence. Polyphone disambiguation is commonly formulated as a classification problem. Therefore a classifier is need to be learned for each character to predict its correct pronunciation in given context. For polyphone disambiguation, machine learning methods like ME [31], CART [29] or MLP are also common used. and the traditional linguistic features (like character, POS, word-terminal syllables etc.) and word embeddings or LM representations are common used as input features.

2.3. Prosodic structure prediction

As shown in Figure 1, a typical prosodic structure of Mandarin TTS system is usually divided into three levels, including the prosodic word (PW), prosodic phrase (PP) and intonation phrase (IP). These three levels of prosody reflect three different length of pause in natural speech signal. Prosodic structure prediction tasks are usually formulated as sequence labeling problems [27]. A sequence of prosodic boundary labels of break (B) or no break (NB) should be predicted for each character or word in input text sequence. Different from other sequence tagging tasks, such as CWS, POS, NER etc., the prosodic structure prediction of each prosodic level predicts the prosodic boundary labels under the constraint of corresponding lower-level labels, e.g. the sequence of prosodic phrases depends on the sequence of prosodic words as indicated in Figure 1. Prosodic structure prediction has been studied over the years, many methods been proposed, including traditional statistical methods, such as ME [11], CRF [14] with manually designed linguistic features and recently proposed RNN-based sequence models [16, 26, 32]. Among these methods, the best reported results were achieved by a BLSTM-CRF [26] based model using text representation extracted by a pre-trained an unidirectional language model.

3. Proposed method using pre-trained text representations

3.1. Transformer based text representations

The main problem of TTS front-end tasks is that there is very limited amount of training data for each task. Although neural network based method has been proposed, most of the conventional approaches have poor generalization on data from new domains since they can not learn the knowledge of a language from such a small dataset. Unsupervised language representation learning therefore became a hotspot in the research area of NLP in recent years. Word embeddings and sentence representations by RNN based language model were utilized as general language knowledge to improve the performance of front-end modules [26]. However, these representations are weak context representations. Word embeddings are only representation of unique words without sequential dependencies. On the other hand, the unidirectional dependency of RNN based language models make it difficult to extract enough context information from the word sequence. Besides, the long-term dependency [22] of a sentence can not be well captured by the recurrent architecture of a language model.

In this work, we proposed to use two kinds pre-training methods to improve the performance of the TTS front-end:

- BERT [23] is a recently proposed unsupervised pre-training method for general NLP tasks. BERT is essentially a language model that can predict words which has been masked out in the input word sequence. It has been reported that the performance of a wide range of NLP tasks can be improved by using word representation extracted by BERT.
- NMT is a well studied NLP task which can commonly access large amount of training data. There are two main components in a typical NMT model: an encoder and a decoder. The encoder takes a sequence of words in source language as input and produces a sequence of context sensitive dense word vectors, based on which the decoder will then output a sequence of words in target language using an attention mechanism.

The transformer [24] architecture is used in both BERT and NMT encoder. The multi-head self-attention mechanism enables the model to capture word dependencies on both left and right side context without any restriction on the position of words in a sentence. On the other hand, a transformer architecture can be easily scale to deep structure and a larger training corpus.

3.2. Application to TTS front-end

Figure 2 shows the architecture of our proposed method using pre-trained text representations for the front-end. The input Mandarin text is firstly pre-processed by TN and CWS modules. Word representations are then extracted by BERT and NMT encoder. These word representations are sent to task specific models. In this work, we focused on the prosodic structure predic-

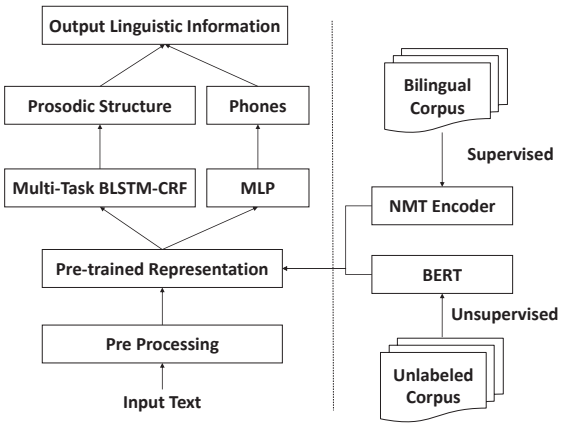

Figure 2: *The architecture of the proposed method using pre-trained text representation for the front-end*

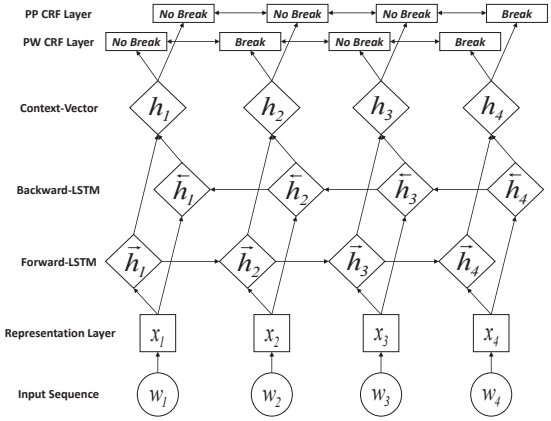

Figure 3: *Architecture of our BLSTM-CRF based prosodic structure prediction model. There are two CRF layers, one for PW, another for PP*

tion and polyphone disambiguation. TN can be accomplished using a method similar with that of polyphone disambiguation, therefore it is not discussed in this paper.

3.2.1. Polyphone disambiguation

A straightforward idea for polyphone disambiguation is to use a separate feedforward neural network as a classifier for each character. However, for most of the characters, the number of training samples is less than 1000. In order to address this data sparsity problem, we use a single compact model for all characters. The categories of all characters are concatenated in output layer and the other layers of all characters share the same parameters. At inference time, we masked the probabilities of pronunciations of irrelevant characters in output layer.

3.2.2. BLSTM-CRF based prosodic structure prediction

The RNN-CRF based model architecture has been successfully applied to many sequence tagging problems. The application of bidirectional LSTM based RNN-CRF (BLSTM-CRF) in prosodic boundary prediction has been studied as a sequence tagging task [27, 33, 26]. The detailed architecture of the model used in this work is shown in Figure 3. Firstly, we use a pre-trained text encoder to encode a sentence into a sequence of word/character vectors. A BLSTM-RNN layer then takes these vectors as input to further produce the input vectors for the CRF layer. We can use two independent models for predicting PW and PP respectively. In addition, since we are using character/word level text representations in an end-to-end way, we can simultaneously predict PP and PW with a single model by a multi-task training.

4. Experiments

4.1. Dataset

We evaluated the proposed representation based methods on Mandarin dataset. We collected the training corpus for polyphone disambiguation and prosodic structure prediction by linguistic experts for our experiments.

For experiments on polyphone disambiguation, we collected a dataset of 300000 sentences. Only one polyphone character was labeled in each sentence. We collected sentences for 89 frequently used polyphone characters, and there are totally 202 character-pronunciation pairs in our corpus. At least 500 samples were collected for each pronunciation of each character. We split the dataset into three subset of 240000, 30000 and 30000 sentences for training, validation and test respectively.

For the experiments on prosodic structure prediction tasks, we collected a corpus of 150000 sentences, which was also split into three subset of 120000, 15000 and 15000 sentences for training, validation and test respectively.

4.2. Experimental setting

In our experiments, we compared several popular text representation methods on our tasks. The BERT model we used in our experiments is pre-trained and released by Google [1]. It has a structure of 12-layer transformer with 768 hidden units in each hidden layer. This character-level model was pre-trained on the Chinese Wikepedia Corpus.

The NMT encoder we used in our experiments was pre-trained on a bilingual corpus with more than 100 million Chinese-English sentence pairs. The NMT model was trained to translate Chinese sentences into English sentences. We adopted a 6-layer transformer as the encoder, and there are 512 hidden units in each of the hidden layers. The encoder takes as input a sequence of subwords in source language [34].

In order to compare with previous representation model, we also trained a character-level bidirectional LSTM based language model [35]. This model was trained on a 50GB of Chinese news text corpus collected from the Internet.

We use the state vector of the last hidden layer as the character representation for all the deep learning based representation models and built the following systems for the experiments in this paper:

1. **ME**: The conventional ME based classifier for polyphone disambiguation using designed linguistic features, including POS tags, the length of words, word combinations left and right context etc [13].
2. **CRF**: the conventional graphical model based sequence labeling method utilizing linguistic features similar with those of ME. The CRF models were applied to PW and PP prediction as a baseline model in our experiments.

[1] https://github.com/google-research/bert

3. **BLM**: the method takes the character representation extracted by character-level bidirectional language model as input.
4. **BERT**: the method takes the character representation extracted by pre-trained BERT as input.
5. **NMT**: the method takes the character representation extracted by pre-trained encoder of an NMT model as input.
6. **TB**: the method that uses feature ensemble by concatenating features of **BERT** and **NMT**.
7. **BERT-MT**: a multi-task method that simultaneously predict PW and PP boundaries in one single model.

4.3. Results and analysis

4.3.1. Evaluation of polyphone disambiguation

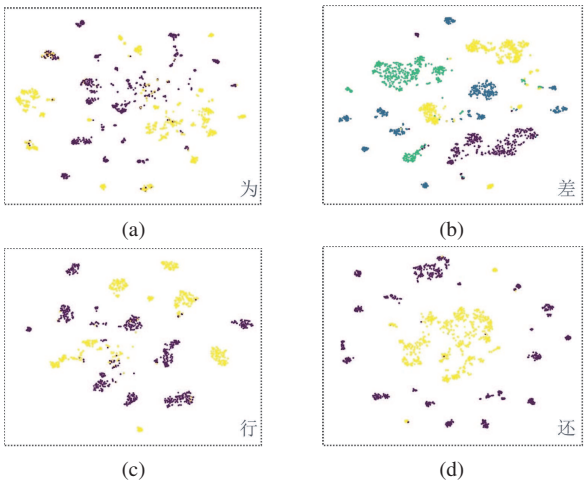

Figure 4: *The t-SNE visualizations of the character representations of four frequently used polyphone characters extracted by BERT. Different colors indicate different pronunciation of a character.*

The BERT is trained to predict randomly missing words/characters in given sentences. Therefore, the model has to learn rich semantic information of each word/character in given context. In order to analyze the space of extracted text representations, we performed a t-SNE [36] analysis on the representation vectors of several polyphone characters on the training set. Figure 4 presents the t-SNE results of four frequently used polyphone characters. One can see that there are clear patterns between different pronunciations of the a character. This make it very easy for the shallow feedforward neural network based classifier to predict correct pronunciation given the character representation. Similar patterns were also observed in the representation space of an NMT encoder.

For the polyphone disambiguation task, we compared the results of proposed method with two baseline systems of ME and BLM. The accuracy rates of the compared systems are presented in Table 1. The results show that transformer based systems significantly outperform the two baseline systems.

By comparing the three transformer based systems, we can see that BERT achieved higher accuracy than NMT encoder in our results, maybe benefit from its deeper structure.

Table 1: *Accuracy rate for different systems in Mandarin on the polyphone disambiguation task*

system	ME	BLM	BERT	NMT	TB
Accuracy	91.34	94.50	96.80	96.18	**96.94**

4.3.2. Evaluation of prosodic structure prediction

For the prosodic boundary prediction task, we compare the results of system NMT with two baseline systems of CRF and BLM. System BLM serves as a strong baseline system here as it achieved best performance in previous researches. As one can see from the results presented in Table 2 that the transformer based methods significantly outperforms these two baseline systems on the both tasks of predicting PW and PP boundaries. However, there are no significant difference between these three transformer based methods, which means that the NMT encoder can achieve a similar performance with a much smaller network structure.

Table 2: *The results of F1 scores of different systems on PW and PP tasks.*

	prosodic word	prosodic phrase
CRF	0.942	0.810
BLM	0.961	0.824
BERT	**0.974**	0.850
NMT	0.974	0.847
BERT-MT	0.973	**0.851**

On the other hand, as the transformer based method can extract rich context sensitive information from the input sequence, it is quite easy for BERT-MT to predict both PW and PP labels using a single model with multi-task training. This can reduce the computational cost of inference at run time. It is difficult to apply the multi-task training to conventional RNN-CRF based methods. Because in the conventional prediction pipeline, PP prediction usually depends on the output of PW prediction.

5. Conclusions

In this paper, we presented an effective method to improve the performance of TTS front-end text processing using pre-trained text representations. Two kinds of transformer based text encoder were investigated in this paper. One is the BERT learned in an unsupervised way and the other is the encoder of an NMT model that trained on bilingual corpus. The experimental results on polyphone disambiguation and prosodic structure prediction show that the proposed methods significantly improved the performance comparing with conventional methods as well as other text representation based methods.

In the future, we will apply the pre-trained transformer text representations to the other modules of TTS front-end, including the text normalization, speaking style prediction, etc.

6. Acknowledgement

The authors would like to thank Chao Bian from smart translation team for fruitful discussions and preparing neural machine translation model for the experiments.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[4] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[5] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *arXiv preprint arXiv:1811.00002*, 2018.

[6] P. Ebden and R. Sproat, "The kestrel tts text normalization system," *Natural Language Engineering*, vol. 21, no. 3, pp. 333–353, 2015.

[7] R. Sproat and T. Emerson, "The first international chinese word segmentation bakeoff," in *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, 2003, pp. 133–143.

[8] L. Màrquez and H. Rodríguez, "Part-of-speech tagging using decision trees," in *European Conference on Machine Learning*. Springer, 1998, pp. 25–36.

[9] 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. Citeseer, 2001, pp. 30–1.

[10] Q. Shi, X. Ma, W. Zhu, W. Zhang, and L. Shen, "Statistic prosody structure prediction," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.* IEEE, 2002, pp. 155–158.

[11] J.-F. Li, G.-p. Hu, and R. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[12] F. Liu, H. Jia, and J. Tao, "A maximum entropy based hierarchical model for automatic prosodic boundary labeling in mandarin," in *2008 6th International Symposium on Chinese Spoken Language Processing*. IEEE, 2008, pp. 1–4.

[13] F. Z. Liu and Y. Zhou, "Polyphone disambiguation based on maximum entropy model in mandarin grapheme-to-phoneme conversion," in *Key Engineering Materials*, vol. 480. Trans Tech Publ, 2011, pp. 1043–1048.

[14] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 135–138.

[15] X. Shen and B. Xu, "A cart-based hierarchical stochastic model for prosodic phrasing in chinese," in *Proc. of ISCSLP00*, 2000, pp. 105–108.

[16] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for chinese speech synthesis using blstm-rnn and embedding features," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 98–102.

[17] 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, 2016, pp. 1–5.

[18] R. Sproat and N. Jaitly, "An rnn model of text normalization." in *INTERSPEECH*, 2017, pp. 754–758.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[22] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies in narx recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 7, no. 6, pp. 1329–1338, 1996.

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

[24] 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*, 2017, pp. 5998–6008.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[26] Y. Zheng, J. Tao, Z. Wen, and Y. Li, "Blstm-crf based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end," *Proc. Interspeech 2018*, pp. 47–51, 2018.

[27] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[28] Y. Huang, Z. Wu, R. Li, H. Meng, and L. Cai, "Multi-task learning for prosodic structure generation using blstm rnn with structured output layer." in *INTERSPEECH*, 2017, pp. 779–783.

[29] F. M. H. G. W. Renhua, "Multi-level polyphone disambiguation for mandarin grapheme-phoneme conversion," *Computer Engineering and Applications*, no. 2, p. 49, 2006.

[30] K. Yao and G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," *arXiv preprint arXiv:1506.00196*, 2015.

[31] X. Mao, Y. Dong, J. Han, D. Huang, and H. Wang, "Inequality maximum entropy classifier with character features for polyphone disambiguation in mandarin tts systems," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4. IEEE, 2007, pp. IV–705.

[32] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5655–5659.

[33] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.

[34] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[35] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," *arXiv preprint arXiv:1705.00108*, 2017.

[36] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.