LEARNING CROSS-LINGUAL KNOWLEDGE WITH MULTILINGUAL BLSTM FOR EMPHASIS DETECTION WITH LIMITED TRAINING DATA

Yishuang Ning[1,2], *Zhiyong Wu*[1,2,3], *Runnan Li*[1,2], *Jia Jia*[1,2,*], *Mingxing Xu*[1,2], *Helen Meng*[3], *Lianhong Cai*[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology (TNList),
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
{ningys13,lirn15}@mails.tsinghua.edu.cn
{zywu,hmmeng}@se.cuhk.edu.hk, {jjia,xumx,clh-dcs}@mail.tsinghua.edu.cn

ABSTRACT

Bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) has achieved state-of-the-art performance in many sequence processing problems given its capability in capturing contextual information. However, for languages with limited amount of training data, it is still difficult to obtain a high quality BLSTM model for emphasis detection, the aim of which is to recognize the emphasized speech segments from natural speech. To address this problem, in this paper, we propose a multilingual BLSTM (MTL-BLSTM) model where the hidden layers are shared across different languages while the softmax output layer is language-dependent. The MTL-BLSTM can learn cross-lingual knowledge and transfer this knowledge to both languages to improve the emphasis detection performance. Experimental results demonstrate our method can outperform the comparison methods over 2-15.6% and 2.9-15.4% on the English corpus and Mandarin corpus in terms of relative F1-measure, respectively.

Index Terms— emphasis detection, cross-lingual, multilingual, bidirectional long short-term memory (BLSTM)

1. INTRODUCTION

Emphasis detection aims to perceive or recognize the emphasized speech segments that may correspond to a word or part of a word from natural speech. As an important prosodic feature, emphasis of speech is not only useful for expressing speakers' emotions and attitudes, but also meaningful for understanding their intentions. Recently, the study of automatic emphasis detection has become an emerging topic that attracts increasing research interests from researchers in speech signal processing. Automatic emphasis detection plays an important role in human-computer interaction scenarios, such as emphatic speech synthesis, content spotting and user intention understanding.

Previous attempts on emphasis detection propose the research problems from two perspectives including features and models. The former mainly focuses on utilizing emphasis related acoustic features for automatic pitch accents recognition and prosodic event detection. For example, [1] used filtered energy features to detect pitch accents. [2] calculated the F0 difference between original speech and synthetic speech, and then used a pre-defined threshold to label emphasis. [3] used spectral emphasis or RASTA-PLP to predict word prominence in spontaneous speech. The latter dedicates to fulfilling the emphasis detection task from the model perspectives. For instance, [4] devised a sound pattern matching (SPM) method for automatic prosodic event detection. [5] considered this task as a classification problem. Motivated by this, [6] proposed to use Bayesian network (BN) with the combination of global acoustic features (F0, duration and energy, semitone) and local acoustic features (tilt parameters). Although the last method can achieve the best performance, the use of tilt parameters involves manual annotation of intonational events; furthermore, just like other classifiers, BN cannot incorporate the contextual information that emphasis detection mainly relies on.

Recently, bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) has shown great potential for leveraging contextual information from both forward and backward directions, and achieved state-of-the-art performance in many sequence processing problems, including voice conversion [7] and speech synthesis [8]. Therefore, it is very suitable for dealing with time sequences like speech for emphasis detection. But the biggest problem is that it needs moderate or large corpus to train a good model [9]. Given that annotating training data is often at great expense, the available training data is always very limited, especially for

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

low-resource language of a specific speaker.

To handle the lacking training data problem, in this paper, we propose a multilingual BLSTM (MTL-BLSTM) model, in which the hidden layers are shared across different languages while the softmax output layers are language-dependent and the input feature vectors of different languages are combined together to form a uniform representation. The shared hidden layers are considered as a universal feature transformation that works well for many languages. The separate softmax layers are used to output the posterior probability of emphasis for each language. Experimental results indicate MTL-BLSTM can learn cross-lingual knowledge and transfer the knowledge between languages to improve the performance of both languages.

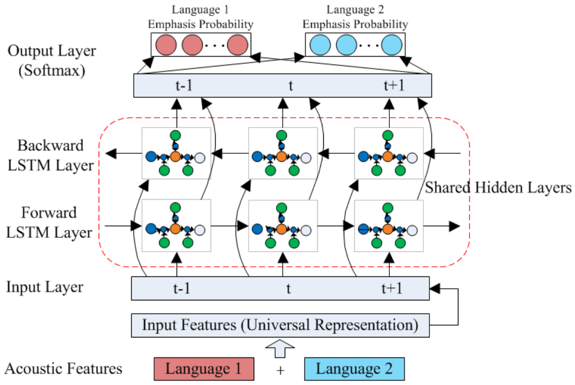

Fig. 1. Architecture of multilingual BLSTM.

2. BIDIRECTIONAL LONG SHORT-TERM MEMORY (BLSTM)

BLSTM-RNN is an extended architecture of bidirectional recurrent neural network (BRNN) [10]. It replaces units in the hidden layers of BRNN with LSTM memory blocks. With these memory blocks, BLSTM can store information for long and short time lags, and leverage relevant contextual dependencies from both forward and backward directions for classification tasks.

BLSTM contains a forward and a backward layer, thus, it can utilize the past and future information for modeling. Given an input sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$, BLSTM computes the forward hidden sequence $\overrightarrow{h}$, the backward hidden sequence $\overleftarrow{h}$ by iterating the forward layer from *t* = 1 to *T*, the backward layer from *t* = *T* to 1:

$$\overrightarrow{h}_t = \phi(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \quad (1)$$

$$\overleftarrow{h}_t = \phi(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \quad (2)$$

The output layer is connected to both forward and backward layers, thus the output sequence can be written as:

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \quad (3)$$

The notations of these equations are explained in [8] and $\phi(\cdot)$ is the activation function which can be implemented by LSTM block with equations in [8].

3. APPROACHES

3.1. Motivation

For the same speech segments, when they are in different contexts, the probabilities of being emphasized are different. Studies reveal that the acoustic features of the emphatic speech are affected by the location of emphasis [11]. Moreover, previous work indicates that emphasis has the characteristic of local prominence and the syllables whose acoustic features (F0, duration, energy) are higher than their neighbors are easier to be perceived as emphasis [11]. That means emphasis is not only closely related with its past acoustic contexts, but also affected by its future ones. Considering both of the past and future acoustic contexts is beneficial for emphasis perception. Motivated by this, we formulate the emphasis detection problem as a sequential learning task and use BLSTM that has tremendous success for leveraging bidirectional contextual information for modeling.

Furthermore, for emphasis perception and detection, there are indeed many intrinsic features that can be shared across different languages. For example, it has been stated for many languages that F0 and duration vary with vowel height, and are constrained by the place of articulation and affected by some contextual aspects [12]. High vowels are consistently shorter than low ones while revealing higher F0. Whether a pitch accent language such as English and Japanese, or a tone language such as Mandarin have shown that the prosodic focus (emphasis) can be realized by F0 variations that are independent of those due to lexical contrasts [13]. Motivated by this, we propose an MTL-BLSTM model which uses the acoustic features as the input. With the shared hidden layers, MTL-BLSTM can learn the cross-lingual knowledge and transfer this knowledge to other languages to improve the performance of emphasis detection.

3.2. Network architecture of MTL-BLSTM

Fig. 1 depicts the architecture of the proposed MTL-BLSTM for emphasis detection of different languages. In this architecture, the input layer and hidden layers are shared across different languages while the output layer is language-dependent. The shared hidden layers (both of the forward and backward layers) can be considered as a universal feature transformation which transforms the input acoustic features of different languages into a language-independent representation. And such representation can be shared across different languages. The output layer is actually a softmax layer and each language has its own softmax layer to estimate the posterior probabilities of the emphasis categories specific to that language.

The main characteristic of MTL-BLSTM is to train the model for both languages simultaneously. However, since the input features are from different languages, the features corresponding to each language are to be used for emphasis detection of its own. To address this problem, the input feature vectors of different languages are combined together to form a single universal feature representation. The dimension of the universal feature representation equals to the sum of the input feature dimensions of language 1 and language 2. For example, when the current input features are from language 1, the universal feature representation is constructed by appending padding symbols (any number not equal to the feature values is OK, i.e. 20 in this paper) behind the features of language 1; When the current input features are from language 2, the universal feature representation is constructed by prepending the same padding symbols before the features of language 1. Then the feature vectors of both languages are shuffled randomly. Finally, the universal feature representation is used as the input of the MTL-BLSTM and transformed by the shared hidden layers to provide benefits to both languages.

3.3. Training procedure

As we can see, MTL-BLSTM is trained with speech data from different languages. This strategy is a variation of multi-task learning (MTL): the tasks of both languages are trained simultaneously. To learn the parameters of our model, we use the mini-batch-based adaptive gradient (Adagrad) [14] algorithm. In each iteration, a task *t* is selected randomly, and the model is updated according to the task-specific objective function. This is actually to minimize the sum of two single-task objectives. We use the softmax loss as the objective functions. For emphasis classification of class c_1 of language 1, the loss function is:

$$loss(c_1, z) = -ln\left(\frac{e^{z_{c_1}}}{\sum_{j=1}^{m} e^{z_j}}\right) = ln\left(\sum_{j=1}^{m} e^{z_j}\right) - z_{c_1} \quad (4)$$

where $c_1 \in \{0,1\}$ is the emphasis label, z_j is the linear prediction of the *j*th category and the loss is summed over all the samples in the mini-batch. The loss function for emphasis detection of language 2 is similar with the one of language 1.

Then we use the trained MTL-BLSTM to detection emphasis of any language used in the training process. By sharing the hidden layers in the model, we can improve the detection performance of both languages.

4. EXPERIMENTS

4.1. Experimental setup

Data set. To evaluate the effectiveness of cross-lingual model transfer, we used the Mandarin (MAN) corpus as language 1 and the English (ENG) corpus as language 2. The MAN corpus is recorded by various speakers from Sogou Voice Assistant. We randomly selected 2000 utterances from the data set and labeled the emphasis for each utterance. The utterances with wrong transcriptions are removed. Finally, we got 1942 utterances, including speeches and their transcriptions. We invited 3 well-trained human labelers to mark the emphasis of each utterance at syllable level by listening to the speech utterances. To address the inconsistency issue, if the labelers had different opinions, they would have a discussion about the inconsistent parts to reach an agreement. As for the ENG corpus, 350 text prompts are carefully designed. After forced alignment, we got 339 utterances in the end. Each text prompt contains one or more emphatic words. For each text prompt, its corresponding speech utterance is recorded with expressive intonation to place proper emphasis on the emphatic words in the sentence.

Features. Previous works indicate that emphasis usually has higher F0, longer duration and higher energy [15]. Besides, research shows the change of semitone is consistent with the distance of auditory perception. This indicates emphasis may be also closely related with semitone. Therefore, the used acoustic features include, F0 related features (mean, minimum, maximum and range of log F0), energy related features (mean, minimum, maximum and range of energy), duration and semitone, totally 10 dimensions. It should be noted that the acoustic features are calculated at syllable for MAN and phoneme for ENG respectively.

Comparison methods. We compared the performance of emphasis detection with some well-known machine learning methods, including support vector machine (SVM), Bayesian network (BN) and conditional random field (CRF). We also designed three more kinds of LSTM models for comparison in addition to the proposed MTL-BLSTM model: 1) monolingual LSTM (MNL-LSTM) trained with the language dependent corresponding corpora and unidirectional LSTM hidden layers; 2) monolingual BLSTM (MNL-BLSTM) trained with the language dependent corresponding corpora and BLSTM hidden layers; 3) mix-lingual BLSTM (MXL-BLSTM) trained using the mixture of two languages without universal feature representation and BLSTM hidden layers.

Evaluation metrics. In all the experiments, we evaluate the detection performance in terms of Precision, Recall and F1-measure [16]. The two corpora are split by train:val:test=8:1:1, with 100 MAN utterances and 30 ENG utterances used as the test set respectively.

4.2. Experimental results

4.2.1. Influence of bidirectional contextual dependencies

Table 1. lists the performance of emphasis detection on ENG test set of using different comparison methods. From the results (in terms of F1-measure), we can see that the performance of using MNL-LSTM is better than that of using other machine learning methods such as SVM, BN and CRF, indicating contextual dependencies are important. Compared with CRF, LSTM can better leverage these contextual dependencies for modeling. Besides, when both past and future contexts are considered (for MNL-BLSTM), the performance can be further improved, which means bidirectional contextual dependencies are useful for our task.

Table 1. Results of using different comparison methods on ENG test set.

Models	Precision	Recall	F1-measure
SVM	0.656	0.810	0.725
BN	0.784	0.792	0.788
CRF	0.785	0.817	0.800
MNL-LSTM	0.823	0.798	0.810
MNL-BLSTM	0.823	**0.821**	**0.822**

Table 2. Performance comparison between different BLSTM derived methods.

(a) Performance on ENG corpus.

Models	Precision	Recall	F1-measure
MNL-BLSTM	0.823	0.821	0.822
MXL-BLSTM	0.826	0.831	0.829
MTL-BLSTM	**0.844**	**0.833**	**0.838**

(b) Performance on MAN corpus.

Models	Precision	Recall	F1-measure
MNL-BLSTM	0.785	0.815	0.799
MXL-BLSTM	0.804	0.822	0.813
MTL-BLSTM	**0.842**	0.803	**0.822**

4.2.2. *Influence of cross-lingual knowledge*

The results on ENG test set are shown in Table 2 (a). As we can see (in terms of F1-measure), both MXL-BLSTM and MTL-BLSTM outperform MNL-BLSTM, which indicates that the model with shared hidden layers is capable of learning cross-lingual knowledge and can transfer this knowledge to other languages. Moreover, the model with uniform feature representation is better than that of simply mixing the samples of different languages. The results demonstrate using large amount of MAN training data is helpful to improve the performance of limited amount of ENG training data, and vise versa (as can be seen in Table 2 (b)).

4.2.3. *Influence of the complementary data*

In this work, ENG is the target language with limited amount of training data. We would like to know to what extent the benefits provided by the data from complementary language (MAN) can be. Shown in Fig. 2 (a), as the scale of the MAN training data increases, the emphasis detection performance on ENG data achieves consistent improvement. Greater performance boost for target (ENG) can be achieved with relatively small scale of the complementary (MAN) data. The results validate the usefulness of the cross-lingual knowledge for emphasis detection.

4.2.4. *Influence of model architectures*

We further evaluate the sensitivity of key parameters related to model architectures, trying to derive the most optimized ones. The number of LSTM memory blocks per hidden layer affects the model performance. Shown in Fig. 2 (b), as the number of blocks increases, the performance gets better at first and then decreases gradually. It achieves the best performance when the number is 64. Besides, we also try different LSTM hidden layers and find that the model with two LSTM hidden layers achieves the best performance. When the layer number is larger than two, the performance decreases gradually, probably because of the over fitting problem caused by limited training data. Hence we use the model with two hidden LSTM layers and 64 LSTM blocks to compare with the baseline methods.

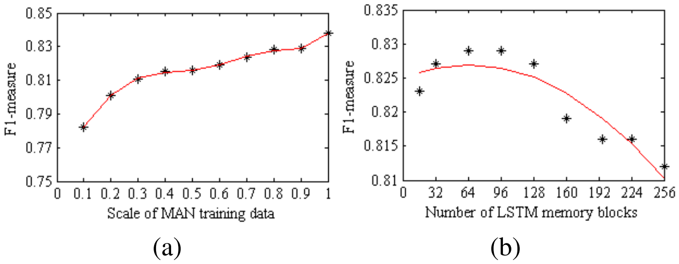

Fig. 2. Emphasis detection performance on ENG test set. (a) Influence of the scale of MAN training data; (b) Influence of the number of LSTM memory blocks per hidden layer.

5. CONCLUSIONS

This paper proposes an MTL-BLSTM model for emphasis detection with limited training data. In this model, the hidden layers are shared across different languages and considered as a universal feature transformation. With this architecture, the cross-lingual knowledge can be learned to provide benefits to both languages. Experimental results demonstrate the effectiveness of our proposed method.

6. ACKNOWLEDGEMENTS

This work is supported by National High Technology Research and Development Program of China (2015AA016305), National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61602033 and 61370023) and joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15). This work is partially supported by major project of National Social Science Foundation of China(13&ZD189).

7. REFERENCES

[1] A. Rosenberg, *Automatic detection and classification of prosodic events*, Ph.D. thesis, Columbia University, 2009.

[2] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based emphatic speech synthesis using unsupervised context labeling," in *Proc. of Annual Conf. of Int. Speech Communication Association (INTERSPEECH)*, 2011, pp. 1849–1852.

[3] A. Schnall and M. Heckmann, "Integrating sequence information in the audio-visual detection of word prominence in a human-machine interaction scenario," in *Proc. of Annual Conf. of Int. Speech Communication Association (INTERSPEECH)*, 2014, pp. 2640–2644.

[4] M. Cernak, A. Asaei, P.E. Honnet, P.N. Garner, and H. Bourlard, "Sound pattern matching for automatic prosodic event detection," Tech. Rep., Idiap, 2016.

[5] F. Tamburini, "Prosodic prominence detection in speech," in *Proc. of IEEE Int. Symposium on Signal Processing and its Applications (ISSPA)*, 2003, pp. 385–388.

[6] Y.S. Ning, Z.Y. Wu, X.Y. Lou, H. Meng, J. Jia, and L.H. Cai, "Using tilt for automatic emphasis detection with Bayesian networks," in *Proc. of Annual Conf. of Int. Speech Communication Association (INTERSPEECH)*, 2015.

[7] L.F. Sun, S.Y. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4869–4873.

[8] Q.J. Yu, P. Liu, Z.Y. Wu, S.Y. Kang, H. Meng, and L.H. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 5545–5549.

[9] Y.C. Fan, Y. Qian, F.L. Xie, and F.K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. of Annual Conf. of Int. Speech Communication Association (INTERSPEECH)*, 2014, pp. 1964–1968.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proc. of IEEE Int. Joint Conf. on Neural Networks*, 2005, pp. 2047–2052.

[11] F.B. Meng, H. Meng, Z.Y. Wu, and L.H. Cai, "Synthesizing expressive speech to convey focus using a perturbation model for computer-aided pronunciation training," in *Proc. of Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, 2010, pp. 22–27.

[12] F. Costa, "Intrinsic prosodic properties of stressed vowels in European portuguese," in *Proc. of Int. Conf. on Speech Prosody*, 2004.

[13] S.W. Chen, B. Wang, and Y. Xu, "Closely related languages, different ways of realizing focus," in *Proc. of Annual Conf. of Int. Speech Communication Association (INTERSPEECH)*, 2009, pp. 1007–1010.

[14] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[15] J.Y. Chen and L. Wang, "Automatic lexical stress detection for chinese learners' of English," in *Proc. of the 7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2010, pp. 407–411.

[16] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," in *Proc. of IEEE Workshop on Spoken Language Understanding (SLU)*, 2012, pp. 210–215.