QUESTION DETECTION FROM ACOUSTIC FEATURES USING RECURRENT NEURAL NETWORK WITH GATED RECURRENT UNIT

Yaodong Tang[1,2]*, Yuchen Huang*[1,2]*, Zhiyong Wu*[1,2,3]*, Helen Meng*[1,3]*, Mingxing Xu*[1,2]*, 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]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

{tangyd14, huang-yc15}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk,
{xumx, clh-dcs}@tsinghua.edu.cn

ABSTRACT

Question detection is of importance for many speech applications. Only parts of the speech utterances can provide useful clues for question detection. Previous work of question detection using acoustic features in Mandarin conversation is weak in capturing such proper time context information, which could be modeled essentially in recurrent neural network (RNN) structure. In this paper, we conduct an investigation on recurrent approaches to cope with this problem. Based on gated recurrent unit (GRU), we build different RNN and bidirectional RNN (BRNN) models to extract efficient features at segment and utterance level. The particular advantage of GRU is it can determine a proper time scale to extract high-level contextual features. Experimental results show that the features extracted within proper time scale make the classifier perform better than the baseline method with pre-designed lexical and acoustic feature set.

Index Terms— question detection, gated recurrent unit (GRU), bidirectional recurrent neural network (BRNN)

1. INTRODUCTION

Detecting questions in human conversations is meaningful for exploring the usage of rich information in speech signal and is an important step in building artificial systems that can better understand natural languages. Question in speech could provide useful clues for identifying speaker's role in a dialog involving multiple speakers [1]. Question detection can be used for automatic meeting index and summarization as question/answer pairs [2].

Previous work on question detection considers not only lexical features but also prosodic-acoustic features. Lexical features have shown better application performances [3][4] and take the key position in question detection system. However, there are two major problems for question detection using lexical features only. First, some questions share the same lexical representation (i.e. words) with its statement form. Second, in most spoken dialog systems, automatic speech recognition (ASR) is the foremost step whose performance will have huge impacts on the following question detection steps [5].

The work in [6] presented the potentials of prosodic feature based classifiers and made it a feasible way to identify English dialog acts from acoustic features. [5][7] extended the work to French and Vietnamese, and found it a useful way to detect interrogative intonation in non-tone languages by using prosodic features with decision tree.

When using classifiers such as decision tree (DT) [3], question detection for tone languages like Mandarin usually begins with extracting features by considering prior phonetic knowledge. However, these features are incomplete for the question detection task. [8] combined prosodic and mel-frequency cepstrum coefficients (MFCCs) for detecting interrogative intonation in Mandarin and achieved improved performance using support vector machine (SVM).

One main problem of previous work is that traditional classifiers like DT or SVM are unable to incorporate the contextual information, on which question detection mainly relies. Recurrent neural networks (RNN) can model context information along time steps of sequence [9]. Inspired by this characteristic, we propose an approach for question detection using improved RNN structures. Specifically, acoustic low level descriptors (LLD) including F0 related features, energy related features and spectral related features are brought in to feed the network with fully integrated information; and then gated recurrent units (GRU) are incorporated for building different RNN structures by taking the advantage of GRU in deciding proper time scale for high-level contextual feature extraction.

The rest of the paper is organized as follows. A brief introduction of used basic RNN structures is given in Section2. Section 3 introduces the proposed models and features in our approach. Experimental setup and results are presented in Section 4. Finally, Section 5 lays out the conclusion and discussion of our work.

2. RECURRENT NEURAL NETWORK

2.1. Bidirectional Recurrent Neural Network (BRNN)

A recurrent neural network (RNN) is able to map from the whole memory of previous inputs to the output at current time step, which is significant for speech signal processing with close time step relationship. Standard RNN can only access information from the previous inputs. However, in speech production, the pronunciation of a segment may also have correlations with the future segments. Bidirectional RNN (BRNN) [10] is able to access past and future context by processing data in both directions.

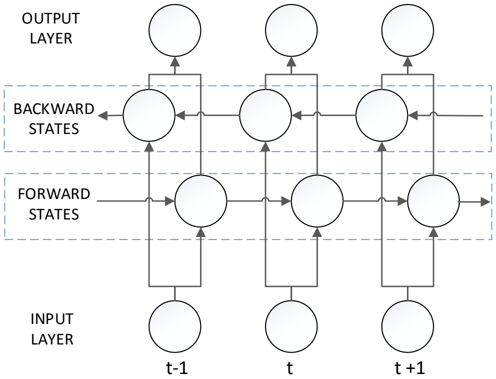


Fig.1. Illustration for the structure of BRNN

As illustrated in Fig.1, the main idea of BRNN is to divide the hidden layer of a standard RNN into forward states part and backward states part. Both parts connect to the same input layer and the same output layer, but without direct connections between the two parts. The difference between the two parts is that forward states are calculated by past inputs along positive time axis while backward states are calculated by future inputs along reverse time axis.

2.2. Gated Recurrent Unit (GRU)

Study of error flow in RNN shows that the standard RNN structure can only keep short-term memory [11] because of the vanishing gradient problem. Gated recurrent unit (GRU) was proposed in [12] to make recurrent blocks adaptively capture the dependencies of different time scales.

Fig.2(b) depicts the illustration of GRU. The activation function (“*f*”, “*g*”) is usually tanh or sigmoid function. In the *j*-th GRU, when given an input vector $\mathbf{x}_t$ at time t, the candidate update $\tilde{h}_t^j$ needs to be calculated first:

$$\tilde{h}_t^j = \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \otimes \mathbf{h}_{t-1}))^j ,$$

where $\otimes$ is an element-wise multiplication operation and a set of reset gates $\mathbf{r}_t$ controls how much the unit updates from all units’ previous activations $\mathbf{h}_{t-1}$ in the same layer. And a reset gate r_t^j is computed by:

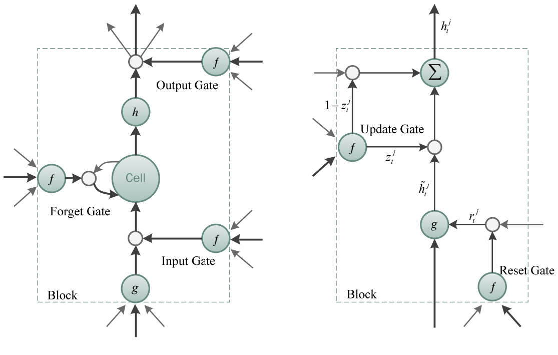
$$r_t^j = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1})^j .$$

The activation function σ is a sigmoid function. Then the activation of the GRU is generated from its previous activation and its current candidate update:

$$h_t^j = (1 - z_t^j)h_{t-1}^j + z_t^j\tilde{h}_t^j ,$$

where the update gate z_t^j decides how much the unit updates from its activation and is calculated as:

$$z_t^j = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1})^j .$$



(a) LSTM (b) GRU

Fig.2. Illustration for the structures of LSTM and GRU

The main difference between GRU and long short term memory (LSTM) [13] is that the memory cell in LSTM does not exist in GRU. As shown in Fig.2(a), there are three gates and one memory cell in a LSTM block. Input gate, forget gate and output gate separately control the data flow from input, memory and output. The activations of input gate, forget gate and output gate depend on current input, previous memory and previous or current output.

While in GRU, the number of gates is reduced. The activations of gates in GRU only depends on current input and previous output. Due to the reduction in parameters, models using GRU tends to converge faster, and the final solution tends to be better than models using LSTM in some tasks [20].

3. APPROACHES FOR QUESTION DETECTION

3.1. Framework

Fig.3 depicts the framework of the proposed approach for question detection. The acoustic feature sequence is first extracted from speech signal of input sentence and normalized at sentence level by sequence standardize module. The *j*-th feature $\mathbf{s}_j^i$ in the *i*-th sequence is standardized to $\mathbf{x}_j^i$ by:

$$\mathbf{x}_j^i = (\mathbf{s}_j^i - \mu_j^i)/\sigma_j^i ,$$

where μ_j^i is the mean of the features and σ_j^i is the standard deviation of the features of the *i*-th feature sequence.

The standardized feature sequence is then fed into the RNN model for extracting high-level contextual features $\mathbf{h}_T$. Collect layer accepts these features for final decision. For the *i*-th sequence, the activation of collect layer is then:

$$c^i = \sigma(\mathbf{W}_c \mathbf{h}_T^i + \mathbf{b}_c) \ ,$$

where $\mathbf{W}_c$ is the weight vector and $\mathbf{b}_c$ is the bias value of collect layer.

For binary classification of our task, a round function is used as the final decider. The label l_i of the *i*-th sequence is assigned by:

$$l_i = round(c^i) \ .$$

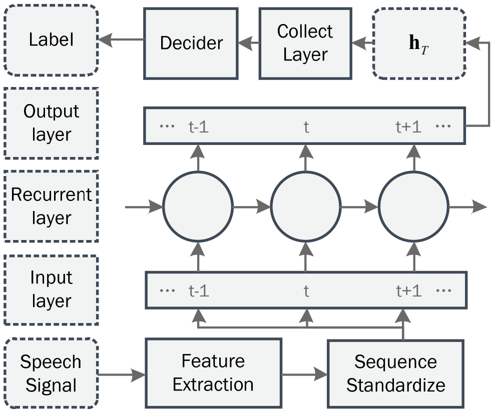


Fig.3. Framework of proposed question detection approach

3.2. Models

To capture different context structures in speech signal for question detection, different models are proposed including GRU-RNN, GRU-BRNN and GRU-DBRNN, serving as the RNN model in the above framework to extract high-level contextual features $\mathbf{h}_T$.

3.2.1 GRU-RNN

We begin with the standard GRU-RNN network structure by replacing the hidden units of RNN with GRUs. For a GRU-RNN with *m* GRUs and an input sequence [$\mathbf{x}_1$,$\mathbf{x}_2$,…, $\mathbf{x}_T$] with time length *T*, the output sequence [$\mathbf{h}_1$,$\mathbf{h}_2$,…,$\mathbf{h}_T$] is calculated by:

$$\mathbf{h}_t = [\Theta_1(\mathbf{x}_t), \Theta_2(\mathbf{x}_t), ..., \Theta_m(\mathbf{x}_t)]^{\mathrm{T}} \ ,$$

where Θ is the activation of a single GRU as in Section 2.

3.2.2 GRU-BRNN

To extract contextual information from both directions, we embed GRU in both forward and backward states of BRNN. For a GRU-BRNN with *m* GRUs in both states and an input sequence [$\mathbf{x}_1$,$\mathbf{x}_2$,…,$\mathbf{x}_T$], the output sequence [$\mathbf{h}_1$,$\mathbf{h}_2$,…,$\mathbf{h}_T$] is a concatenation from both states:

$$\mathbf{h}_t = [\Theta_1(\mathbf{x}_t), ..., \Theta_m(\mathbf{x}_t), \Phi_1(\mathbf{x}_t), ..., \Phi_m(\mathbf{x}_t)]^{\mathrm{T}} \ ,$$

where Θ is the activation of a single GRU in forward states with memory from 1 to *t*-1, Φ is the activation of a single GRU in backward states with memory from *T* to *t*+1. Both of them share the same calculation method as in Section 2.

3.2.3 GRU-DBRNN

Aiming at modeling higher-level representation, we build GRU-DBRNN model by stacking two GRU-BRNN layers. The first GRU-BRNN layer with *m* GRUs in both states receives the input sequence [$\mathbf{x}_1$,$\mathbf{x}_2$,…,$\mathbf{x}_T$] and derives the output [$\mathbf{h}_1$,$\mathbf{h}_2$,…,$\mathbf{h}_T$]^1 by:

$$\mathbf{h}_t^1 = [\Theta_1^1(\mathbf{x}_t), ..., \Theta_m^1(\mathbf{x}_t), \Phi_1^1(\mathbf{x}_t), ..., \Phi_m^1(\mathbf{x}_t)]^{\mathrm{T}} \ ,$$

where Θ^1 and Φ^1 are the activations of a single GRU in forward states and backward states of the first layer.

Then the second GRU-BRNN layer generates the final output sequence [$\mathbf{h}_1$,$\mathbf{h}_2$,…,$\mathbf{h}_T$] by:

$$\mathbf{h}_t = [\Theta_1^2(\mathbf{h}_t^1), ..., \Theta_m^2(\mathbf{h}_t^1), \Phi_1^2(\mathbf{h}_t^1), ..., \Phi_m^2(\mathbf{h}_t^1)]^{\mathrm{T}} \ ,$$

where the forward activation and backward activation of a single GRU in the second layer are Θ^2 and Φ^2.

3.3. Features

In previous research, there is a universal hypothesis that pitch carries the major information of question. Though typical interrogative intonation in Mandarin questions is formed with the rising boundary tone, there still are many variations due to the sentence type and sentence structure [14]. Inspired by work on speech emotion recognition [15], we use acoustic low level descriptors (LLD) and their first order derivatives as the basic acoustic features.

Covering short-term features in each time frame, LLD can be used to characterize any type of sound [16]. In this work, we use LLD feature set proposed in INTERSPEECH 2014 Computational Paralinguistic Challenge [17] with total 65 features, including 4 energy related LLD (Loudness 2, Energy 1, ZCR 1), 55 spectral related LLD (Bands 26, MFCC 14, Others 15) and 6 voicing related LLD (F_0 2, Prob. of voicing 1, HNR 1, Jitter 2, Shimmer 1).

4. EXPERIMENT

4.1. Experimental Setup

We use a simulated Call Center Recording of Mandarin as our experiment dataset [18]. The simulated corpus consists of 408 dialogs with duration between 40 to 90 seconds. Each dialog involves 2 speakers. 20 native speakers of Mandarin from college students were invited to participate in recoding the dialogs. 3 labelers were asked to annotate each sentence in dialogs with the sentence type manually. If there were any inconsistences between the annotated results, the labelers were gathered together for discussion to reach the agreement.

There are 2,850 question sentences (Q) and 3,142 non-question sentences (NQ) like statements or greetings. Each sentence is saved in Microsoft wav format and sampled at 16 KHz. The 20ms windows size and 10ms window shift are then used to convert the raw wav data into frames. OpenSMILE [19] is chosen as the feature extractor to obtain the LLD (as in Section 3.3) and their first order derivatives with 130 dimensions in total.

Table 1. The number of units in each hidden layer. "F" is the forward states and "B" is the backward states in BRNN.

Model	L1		L2		# Params
GRU-RNN	128		-		132k
LSTM-RNN	128		-		99k
Model	F	B	F	B	# Params
GRU-BRNN	64	64	-	-	75k
LSTM-BRNN	64	64	-	-	99k
GRU-DBRNN	32	32	32	32	50k
LSTM-DBRNN	32	32	32	32	67k
Model	L1	L2	L3	L4	# Params
Simple-DNN	128	128	128	128	66k

Besides the three GRU based neural network models in Section 3, we also implement LSTM models with the same structure as GRU networks and a simple DNN model with 4 hidden layers for performance comparison. All the implementations of the models are based on Theano [21, 22] and Keras 0.1.2 [23]. Table 1 summarizes the network architectures of different models showing the unit number at each layer in each model. The normalized uniform initialization in [24] is used to initialize the hidden layers of the above models, and Adam [25] ($lr = 1e-3, \varepsilon = 1e\text{-}8$) is adopted for training models.

According to the previous work [3], we choose C4.5 decision tree based hybrid system using lexical and acoustic features (DT-LA) as the baseline. The lexical features used in this system include interrogative pronouns, sentence final particle, A not A construction, the positions of the terms, the number of words in whole sentence and the word "you" or "your". Acoustic features are derived from our 130 LLD feature set in Section 3.3 by applying 7 functions including max, min, mean, variance, range, lower and upper quartile. SVM and the above simple DNN model are used to classify sentence type from acoustic features only, where the acoustic feature set is the same as used in DT-LA system.

4.2. Experiment Results

We use the measurements precision (P), recall (R) and F1-measure for objective evaluation as follows:

$$P = \frac{N_{correctly\_detected\_questions}}{N_{total\_detected\_questions}} ,$$

$$R = \frac{N_{correctly\_detected\_questions}}{N_{total\_questions}} ,$$

$$F1 = \frac{2P \cdot R}{P + R} ,$$

where $N_{correctly\_detected\_questions}$ is the number of true questions detected, $N_{total\_detected\_questions}$ is the number of all questions detected, $N_{total\_questions}$ is the number of true questions in our test set. The final results are obtained from the mean values of these measurements in 5-folds cross validation.

Table 2. Objective evaluation results of each model.

Model	Precision	Recall	F1
SVM	68.7%	71.3%	70.0%
DT-LA	73.0%	80.0%	76.4%
Simple-DNN	74.0%	76.5%	75.3%
GRU-RNN	**87.1%**	79.0%	82.8%
LSTM-RNN	80.5%	85.4%	82.9%
GRU-BRNN	83.6%	**87.5%**	**85.5%**
LSTM-BRNN	82.6%	83.8%	83.2%
GRU-DBRNN	82.8%	84.9%	83.9%
LSTM-DBRNN	84.0%	82.3%	83.1%

From the F1-measure in Table 2, when using acoustic features only for classification, our RNN based approaches outperform the SVM method and the simple DNN model. Compared to the simple DNN model, contextual features extracted by RNN structures are efficient for the question detection task. In our experiment, GRU based networks tend to get slightly better result than LSTM RNNs. At least 6% absolute improvement in F-measure than the DT-LA system demonstrates RNNs' usefulness in modelling particular context information in acoustic aspect.

5. CONCLUSION

In this work, we proposed gated recurrent unit (GRU) based recurrent network for the task of detecting questions in Mandarin conversational speech. GRU can be used to model context information of speech question at segment and utterance level from acoustic features. We do not have to use pre-designed or selected feature set, and can get better result than previous work with combined lexical features and utterance level acoustic features. Results indicate that, in complex network structures, GRU with less parameters tends to perform slightly better than LSTM. We will concentrate on fine-tuning our models and making analysis on how RNN models context information in our question detection task.

6. ACKNOWLEDGEMENT

This work is supported by the National Basic Research Program of China (2012CB316401), the National High Technology Research and Development Program of China (2015AA016305), the National Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023 and 61171116), the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15) and the Major Program for National Social Science Foundation of China (13&ZD189).

7. REFERENCES

[1] T. Bazillon, B. Maza, M. Rouvier, F. Bechet, A. Nasr, “Speaker role recognition using question detection and characterization,” in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 1333-1336, 2011.

[2] A. Kathol, G. Tur, “Extracting question/answer pairs in multi-party meetings,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5053-5056, 2008.

[3] J. Yuan, D. Jurafsky, “Detection of questions in Chinese conversational speech,” in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, pp. 47-52, 2005.

[4] K. Boakye, B. Favre, D. Hakkani-Tür, “Any questions? Automatic question detection in meetings,” in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, pp. 485-459, 2009.

[5] V.M. Quang, E. Castelli, P.N. Yên, “A decision tree-based method for speech processing: question sentence detection,” in *Fuzzy Systems and Knowledge Discovery*, pp. 1205-1212, 2006

[6] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, C.V. Ess-Dykema “Can prosody aid the automatic classification of dialog acts in conversational speech?” *Language and Speech*, vol. 41, no. 3-4, pp. 439-487, 1998.

[7] V.M. Quang, L. Besacier, E. Castelli, “Automatic question detection: prosodic-lexical features and cross-lingual experiments,” in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 2257-2260, 2007.

[8] W. Bao, Y. Li, M. Gu, J. Tao, L. Chao, S. Liu “Combining prosodic and spectral features for Mandarin intonation recognition,” in *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 497-500, 2014.

[9] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Heidelberg: Springer, 2012.

[10] M. Schuster, K.K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, 1008.

[11] A. Graves, J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5, pp. 602-610, 2005.

[12] K.H. Cho, B.V. Merriënboer, D. Bahdanau, Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv*: 1409.1259, 2014.

[13] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[14] Y. Wang, J. Jia, L. Cai, “Analysis of Chinese interrogative intonation and its synthesis in HMM-Based synthesis system,” in *Proc. International Conference on Internet Computing and Information Services (ICICIS)*, pp. 343-346, 2011.

[15] C.N. Anagnostopoulos, T. Iliou, I. Giannoukos, “Features and classifiers for emotion recognition form speech: a survey from 2000 to 2011,” *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155-177, 2015.

[16] H.G. Kim, N. Moreau, T. Sikora, “MPEG-7 audio and beyond: Audio content indexing and retrieval,” *John Wiley & Sons*, 2006.

[17] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, Y. Zhang, “The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive and physical load,” in *Proc. Annual Conference of International Speech Communication Association (INTER-SPEECH)*, 2014.

[18] A. Li, M. Xu, L. Cai, “Acoustic features prominence based Chinese question detection,” *Chinese Sciencepaper*, vol. 9, no. 7, pp. 826-829, 2014.

[19] F. Eyben, M. Wöllmer, B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proc. International Conference on Multimedia*, pp. 1459-1462, 2010.

[20] J. Chung, C. Gulcehre, K.H. Cho, Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv*:1412.3555, 2014.

[21] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, Y. Bengio, “Theano: new features and speed improvements,” in *Proc. NIPS 2012 deep learning workshop*, 2012.

[22] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, “Theano: A CPU and GPU mjath expression compiler,” in *Proc. Scientific Computing with Pythons (SciPy)*, 2010.

[23] F. Chollet, Keras [OL]. [2015-09-08]. *GitHub repository*, https://github.com/fchollet/keras.

[24] X. Glorot, Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[25] D. Kingma, J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference for Learning Representations (ICLR)*, 2014.