Interspeech 2018
2-6 September 2018, Hyderabad

Automatic question detection from acoustic and phonetic features using feature-wise pre-training

Atsushi Ando[1], *Reine Asakawa*[1,2], *Ryo Masumura*[1], *Hosana Kamiyama*[1], *Satoshi Kobashikawa*[1], *Yushi Aono*[1]

[1] NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2] Toyohashi University of Technology, Japan

{ando.atsushi, masumura.ryo, kamiyama.hosana, kobashikawa.satoshi, aono.yushi}@lab.ntt.co.jp, asakawa@nlp.cs.tut.ac.jp

Abstract

This paper presents a novel question detection method from natural speech using acoustic and phonetic features. The conventional methods based on Recurrent Neural Networks (RNNs) use only acoustic features. However, lexical cues are essential to identify some questions such as declarative questions. To this end we propose a new RNN-based question detection model which utilizes both acoustic and lexical information. Phonetic features which are suitable to describe interrogative cues are used as lexical information. Furthermore, we also propose a new training framework named feature-wise pre-training (FP) to combine the acoustic and phonetic features effectively. FP attempts to acquire interrogative cues in individual features instead of the combination of the features, which makes the model training more stable. The estimation models of the interrogatives are then integrated and fine-tuning is applied to obtain the unified comprehensive model. Experiments show that the proposed method offers better performance than the conventional benchmarks.

Index Terms: Question Detection, Recurrent Neural Networks (RNNs), Senone bottleneck, Feature-wise Pre-training

1. Introduction

Question detection from natural speech is an important technology to understand human intention. It has so many applications such as generating relevant responses in spoken dialog systems [1], meeting indexing [2] and improving machine translation from speech [3]. The aim of this paper is question detection which is regarded as the task of classifying individual utterances into two classes: *question* or *statement*.

A considerable number of studies have examined question detection. According to several researchers [4, 5], cues indicative of questions appear in two aspects: acoustics and lexicon. The most typical acoustic cue is rising pitch in the end of an utterance. Lexical cues are particular words or phrases like interrogative words (*'wh-'*) and second-person pronouns (*'you'*, *'your'*, etc.) [6]. Most of the conventional methods use heuristic features in an attempt to capture these types of information. The statistics of fundamental frequencies (F_0) over both the whole utterance and last 200-500 ms are often used as acoustic features [4,7]. Lexical features generally include n-grams of words or part-of-speech tags [4, 6, 8]. These are based on the outputs of automatic speech recognition systems because it is impossible to use transcriptions in practice. Finally, common two-class classifiers such as decision tree have been utilized to estimate questions from these two types of heuristic features.

One of the problems of these heuristic methods is that fixed features available are insufficient for robust question detection. For example, some questions may exhibit pitch rise for longer durations than the feature analysis intervals. Recently, a new question detection framework based on Recurrent Neural Networks (RNNs) was proposed to solve this problem [9, 10]. It can capture contextual changes from a series of acoustic features without any heuristics. They reported that the RNN-based approach yielded better performance than the methods that use heuristic features in simulated dialogs.

Though lexical cues are essential to identify some questions such as declarative questions, the conventional RNN-based method ignored lexical information. In this paper, we propose a new RNN-based question detection model which utilizes both acoustic and lexical information. To the best of our knowledge, this is the first RNN-based work that deals lexical information for question detection without any transcription. As lexical information, this paper uses frame-wise phonetic features called senone bottleneck features. They have been widely used as an indicator of lexical information in several tasks like spoken language recognition [11, 12] or speaker recognition [13]. We consider these phonetic features to be suitable for lexical information in question detection because lexically interrogative cues often exhibit similar phoneme sequences like *'wh-'*.

Furthermore, we also propose a new training framework for a question detection model with two types of features. Questions have several subtypes; some questions have only acoustic cues, while others have only lexical cues. It follows that the RNN-based model has difficulty in learning these complex relationships from the combination of the features. To solve this problem, we utilize the identification rule of questions. It is considered that utterances are determined as questions if interrogative cues appear in at least one of acoustic and phonetic aspects. This indicates that it is rather important to acquire interrogative cues in individual features than the combinations of them. The proposed training framework, named feature-wise pre-training (FP), attempts to acquire this property by two steps. First, estimation models of acoustic or phonetic interrogatives are trained independently. The estimation models are then integrated and fine-tuning is applied to obtain a unified comprehensive model.

The contributions of this paper are as follows:

- This is the first work on RNN-based question detection to use both acoustic and lexical information without transcriptions. Frame-wise phonetic features are employed as lexical information.
- A new training framework named feature-wise pre-training is proposed to construct robust question detection models with two types of features.

10.21437/Interspeech.2018-1755

Table 1: *Question types and examples.*

		Acoustic	
		interrogatives (pitch rise in the end, ...)	declaratives (pitch fall in the end, ...)
lexical	interrogatives (*yes-no*, *wh-*, ...)	**proper questions** *Have you looked at that?*	**lexical-only questions** *What was the nature of the email?*
	declaratives (single word, ...)	**acoustic-only questions** *Tomorrow?*	**statements** *Please play some music.*

- Utterances gathered from spoken dialog systems in real environments are used for performance evaluation. Some of the conventional methods were tested using only simulated dialogs.

This paper is organized as follows. Question types are discussed to clarify the difficulty of question detection in Section 2. The conventional method that uses RNNs and acoustic features is introduced in Section 3. Section 4 presents the proposed method. Experiments and the results are discussed in Section 5 while our conclusions are given in Section 6.

2. Question types

To clarify the task of question detection, we first categorize the types of questions. According to previous research [4] and our analysis of natural speech, question types can be categorized from acoustic and lexical aspects, see Table 1. Conventional research terms acoustic-only questions as declarative questions. Question detection is formulated as the task of classifying these three types of questions (proper, acoustic-only and lexical-only questions) into *question* class; the remainder are assigned to *statement*. Some studies have shown that proper questions and acoustic-only questions exhibit near identical acoustic cues [5].

Table 1 shows two factors. First, it is essential for both acoustic and lexical information to identify all questions. Second, utterances are determined as questions if interrogative cues appear in at least one of acoustic and lexical aspects.

3. Conventional method

This section describes the conventional question detection method that uses Recurrent Neural Networks (RNNs) with acoustic features [9].

Let $\mathbf{X}^{(a)} = [\mathbf{x}_1^{(a)}, \cdots, \mathbf{x}_{T_a}^{(a)}]$ be the acoustic features of an utterance. The conventional method estimates the correct class c from $\mathbf{X}^{(a)}$:

$$\hat{c} = \underset{c \in \{c_0, c_1\}}{\arg\max}\, P(c \mid \mathbf{X}^{(a)}, \theta_a), \tag{1}$$

where $\hat{c}$ is an estimated class, T_a is the length of the acoustic features, and θ_a is the set of the parameters of the question detection model. c_0 and c_1 represents *statement* and *question*, respectively.

The structure of the classifier of the conventional method is shown in Fig. 1. In this figure, the output of the model $\mathbf{y} = [P(c_0 \mid \mathbf{X}^{(a)}, \theta_a), P(c_1 \mid \mathbf{X}^{(a)}, \theta_a)]$ represents posterior probabilities of the classes. The model parameters θ_a are optimized by the set of the training utterances.

Though the conventional method showed a good performance in simulated utterances, natural speech contains lexical-only questions which is impossible to detect from acoustic features alone. Therefore, lexical information is essential for question detection in real environments.

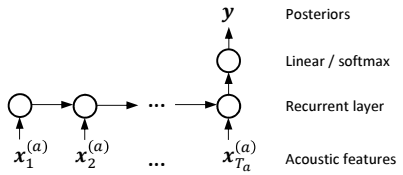

Figure 1: *An example of the structure of the question detection model based on RNNs with acoustic features [9].*

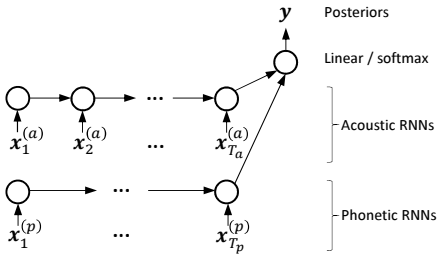

Figure 2: *The structure of the proposed question detection model with acoustic and phonetic features.*

4. Proposed method

4.1. Model structure

In this paper, we propose a new RNN-based question detection approach; it utilizes not only conventional acoustic information but also lexical information.

We use frame-wise phonetic features called senone bottleneck features as lexical information. We consider them to be suitable for lexical information in question detection. One of the reasons is that lexically interrogative cues often exhibit similar phoneme sequences like *'wh-'*. Another is that they will be more robust in speech recognition errors than word-based features because word features are shown completely different characteristics if recognition errors are occurred.

The proposed method estimates the correct class from both acoustic features $\mathbf{X}^{(a)}$ and phonetic features $\mathbf{X}^{(p)} = [\mathbf{x}_1^{(p)}, \cdots, \mathbf{x}_{T_p}^{(p)}]$,

$$\hat{c} = \underset{c \in \{c_0, c_1\}}{\arg\max}\, P(c \mid \mathbf{X}^{(a)}, \mathbf{X}^{(p)}, \Theta), \tag{2}$$

where T_p is the sequence length of the phonetic features and Θ is the set of model parameters of the proposed model. Note that T_a and T_p may be different because the feature extraction periods of acoustic and phonetic features may not be the same. The model structure of the proposed method is shown in Fig. 2.

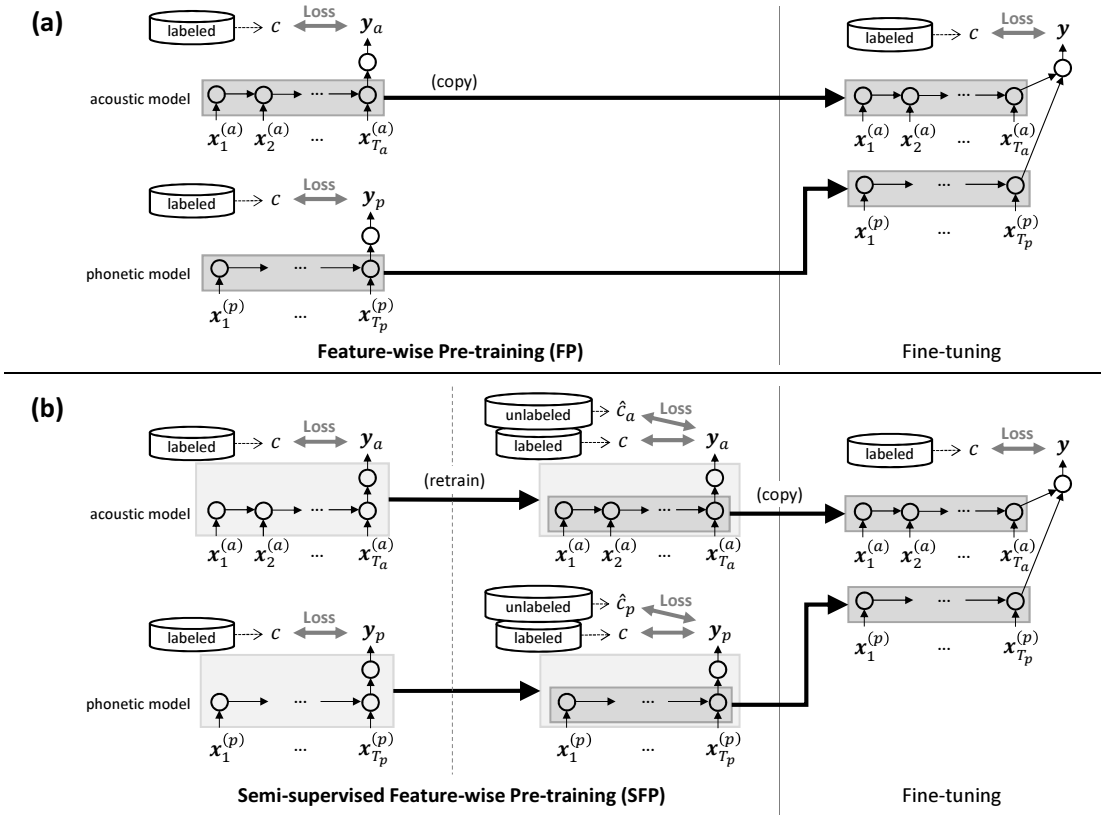

Figure 3: *Flow of the proposed training; (a) Feature-wise Pre-training and (b) Semi-supervised Feature-wise Pre-training.*

4.2. Model training

4.2.1. Feature-wise Pre-training (FP)

Training the proposed classifier raises one problem: the difficulty of stable training. There are several types of questions as noted in Section 2. For example, some questions have acoustic cues but not lexical cues, while others have lexical cues but not acoustic cues. It is difficult to learn these complex relationships from the combination of the features, especially in a small training data. This makes the training of the proposed classifier unstable, which degrades classification performances.

To solve this problem, we propose a new training framework named feature-wise pre-training (FP) for the proposed model. FP utilizes the characteristics of questions described as the second factor in Section 2: questions are determined if interrogative cues appear in at least one of acoustic and lexical aspects. In other words, it is rather important to acquire interrogative cues in individual features than the combinations of the features. FP takes this strategies, i.e. train question detection model from individual features at first, then integrated two models to construct a unified comprehensive model. Each step of the FP is easier to learn than training the complex combinations of the two types of features, which will yield stable training and better performance.

There are two hypotheses in FP. First, individual features are enough to distinguish questions. Second, acoustic-only and lexical-only questions are the minorities of all questions. That is, more than half of questions have interrogative cues in acoustics, and other more than half of questions have interrogative cues in lexicon.

The flow of FP is shown in Fig. 3(a). It has two steps. First, train question estimation models for each feature. In this step, original teacher labels in the training data are used as the ground truth of the individual estimation models. It means that the models are trained as if every questions have interrogative cues in both acoustic and phonetic features. This is permissible if more than half of questions have interrogative cues in each feature, which is the second hypothesis. Second, RNN parts of the individual estimation models are copied as initial parameters of the the unified model. Finally fine-tuning of the unified model yields a comprehensive question detection model.

4.2.2. Semi-supervised Feature-wise Pre-training (SFP)

It is considered that improving individual estimation models of the features yields the better unified model. Therefore, we introduce semi-supervised training in feature-wise pre-training to reinforce individual models. The overview of this approach named semi-supervised feature-wise pre-training (SFP) is shown in Fig. 3(b). Self-training [14], a well-known semi-supervised training method, is used in SFP.

The flow of the SFP is as follows. First, individual models are trained separately with original teacher labels, which is same as the first step of FP. These models are then retrained separately by self-learning. Several non-labeled utterances are selected by evaluating and thresholding the output posterior probabilities of the non-labeled datasets, and used as an additional training data. Finally, the recurrent layers of these retrained models are integrated and fine-tuned to create the unified model.

Table 2: *Overall accuracies and precisions, recalls, F-measures of question class.*

		Features			*question*		
	Training method	Acoustic	Phonetic	Accuracy	Precision	Recall	F-measure
Conventional [9]	Flat	✓		86.3	50.7	52.6	51.6
Proposed	Flat		✓	**90.4**	**69.8**	54.7	61.4
		✓	✓	86.4	51.3	44.3	47.5
	FP + Fine-tune	✓	✓	90.1	65.5	**61.2**	**63.3**
	SFP	✓		86.9	52.7	51.9	52.3
			✓	89.7	63.3	61.2	62.2
	SFP + Fine-tune	✓	✓	**90.5**	**66.8**	**62.8**	**64.7**

5. Experiments

5.1. Setup

We compared the performances of the conventional and the proposed method by 10-fold cross validation evaluation.

For our experiments, we newly collected Japanese speech samples toward a spoken dialog system for real environments. The samples contained web search, asking weather and chat, etc. Speakers ranged from child to elder, and speaker identification was dropped because the system did not permit the input of personal information. The sampling rate of audio signals was 16 kHz. Every sample was subjected to voice activity detection and those that contained no speech were eliminated. Note that some utterances contained background music or speech noise. Average utterance length was 2.4 sec. One annotator gave training labels (*question* or *statement*) to individual utterances. This yielded 1056 questions, 6553 statements and 200 thousand non-labeled utterances. We evaluated reliability of the training labels by sampling and re-annotating 500 utterances, the Kappa coefficient was 0.889 which indicates that we can have confidence in the labels. In cross validation step, we used one subset as test, another one as development and the rest 8 subsets as training set.

We used the same acoustic features as the conventional methods [9, 10]. They are the low-level descriptors proposed in INTERSPEECH 2014 Computational Paralinguistic Challenge [15]. They had 65-dimensional values (4 energy-related, 55 spectral-related and 6 voice-related) and their first order derivatives; this yielded a total of 130 dimensions in every frame. We employed 20 ms window and 10 ms window shift in this experiment. All the acoustic features were extracted by OpenSMILE [16] and were z-normalized for each utterance.

64 dimensional senone bottleneck features were used as phonetic features. The senone recognition model consisted of 384 dimensional Mel-Frequency Cepstral Coefficients (MFCCs) input, four hidden linear layers (1024 hidden units in first three layers and 64 in bottleneck layer) and 3072 dimensional output senone states. The senone model was trained by the Corpus of Spontaneous Japanese (CSJ) [17]. Training involved around 600 hours of speech and senone state accuracy was 69.2 % in the 2 hour evaluation data in the same corpus, which demonstrated adequate performance for phoneme recognition.

In this experiment, we compared the performance of the combinations of the features and several training methods. The training methods included flat-start which use random values as initial weights (flat), fine-tuning after feature-wise pre-training (FP+Fine-tune), semi-supervised feature-wise pre-training (SFP), and fine-tuning after SFP (SFP+Fine-tune). The models of the flat-start with individual features are equivalent to the those of FP. Gated Recurrent Units (GRUs) with 128 hidden units were used in both the conventional and the proposed method. The conventional method was flat-start with acoustic features only [9]. The variants of the proposed method were GRUs with phonetic features alone, and with both acoustic and phonetic features trained by FP or SFP. The optimization algorithm was Adam [18] and dropout ratio was 0.5. Learning rate was 0.001 in the flat-start, 0.0001 in semi-supervised retraining and fine-tuning. Early-stopping was tested by development set in every epoch to prevent over-fitting. The data selection threshold of the semi-supervised retraining in SFP is 0.8 and 0.6 in acoustic and phonetic feature models, respectively. Overall accuracy and precision, recall, F-measure of *question* class were used to compare the performances.

5.2. Results and Discussions

The performances of the conventional and the proposed methods are shown in Table 2. A comparison of the features, see the first and second rows, shows that phonetic features yielded better performance, as the individual differences in lexical cues are smaller than those of acoustic features.

The combination of the acoustic and phonetic features with flat-start yielded worse performance than the use of the individual features, see in the third row. However, FP and fine-tuning yielded better performance than individual features. This indicates that it is difficult to acquire several types of questions from limited data, but FP approach which constructing individual evaluators first helps to achieve accurate classification.

Finally, we compare the proposed feature-wise pre-training with and without semi-supervised training. A comparison of the results of flat-start and SFP of individual features shows that the semi-supervised approach attains higher accuracy. This means semi-supervised training enables to reinforce individual estimation models. Furthermore, the combination of SFP and fine-tuning showed the best performance. These results reveal the effectiveness of the proposed training framework.

6. Conclusions

In this paper, we proposed a new question detection method that uses acoustic and phonetic features. The proposed method employs senone bottleneck features as a compact representation of lexical characteristics. Furthermore, the proposed method uses a new training framework that can acquire several types of questions. The proposed framework trains individual evaluators of questions from acoustic or phonetic features. These evaluators are then integrated and fine-tuned to yield the final comprehensive model. Experiments on Japanese utterance gathered for spoken dialog systems showed the effectiveness of both phonetic features and the proposed training framework in question detection. Future works include evaluation with larger dataset.

7. References

[1] E. Shriberg, R. A. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van 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, 2000.

[2] A. Kathol and G. Tur, "Extracting question/answer pairs in multi-party meetings," in *Proc. of ICASSP*, 2008, pp. 5053–5056.

[3] Y. Wakita, J. Kawai, and H. Iida, "Correct ptarts extraction from speech recognition results using semantic distance calculation, and its application to speech translation," in *Proc. of an ACL Workshop on Spoken Language Translation*, 1997, pp. 24–31.

[4] K. Boakye, B. Favre, and D. Hakkani-Tür, "Any questions? automatic question detection in meetings," in *Proc. of ASRU*, 2009, pp. 485–489.

[5] M. Safanova and M. Swerts, "On recognition of declarative questions in english," in *Proc. of Speech and Prosody*, 2004, pp. 313–316.

[6] J. Liscombe, J. J. Venditti, and J. Hirschberg, "Detecting question-bearing turns in spoken tutorial dialogues," in *Proc. of INTERSPEECH*, 2006, pp. 69–72.

[7] A. Margolis and M. Ostendorf, "Question detection in spoken conversations using textual conversations," in *Proc. of ACL-HLT*, 2011, pp. 118–124.

[8] Q. Vu, L. Besacier, and E. Castelli, "Automatic question detection: prosodic-lexical features and crosslingual experiments." in *Proc. of INTERSPEECH*, 2007, pp. 2257–2260.

[9] Y. Tang, Y. Huang, Z. Wu, H. Meng, M. Xu, and L. Cai, "Question detection from acoustic features using recurrent neural network with gated recurrent unit," in *Proc. of ICASSP*, 2016, pp. 6125–6129.

[10] Y. Tang, Z. Wu, H. M. Meng, M. Xu, and L. Cai, "Analysis on gated recurrent unit based question detection approach," in *Proc. of INTERSPEECH*, 2016, pp. 735–739.

[11] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Study of senone-based deep neural network approaches for spoken language recognition," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 105–116, 2016.

[12] R. Masumura, T. Asami, H. Masataki, and Y. Aono, "Parallel phonetically aware dnns and lstm-rnns for frame-by-frame discriminative modeling of spoken language identification," in *Proc. of ICASSP*, 2017, pp. 5260—5264.

[13] A. Lozano-Diez, A. Silnova, P. Matějka, O. Glembek, O. Plchot, J. Pešán, L. Burget, and J. Gonzalez-Rodrigues, "Analysis and optimization of bottleneck features for speaker recognition," in *Proc. of Odyssey*, 2016, pp. 352–357.

[14] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. of ACL*, 1995, pp. 189–196.

[15] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: Cognitive and physical load," in *Proc. of INTERSPEECH*, 2014, pp. 427–431.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, 2010, pp. 1459–1462.

[17] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *Proc. of SSPR*, 2003, pp. 7–12.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980