TOPIC INDEPENDENT LANGUAGE MODEL FOR KEY-PHRASE DETECTION AND VERIFICATION

Tatsuya Kawahara *Shuji Doshita*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

A topic independent lexical and language modeling for robust key-phrase detection and verification is presented. Instead of assuming a domain specific lexicon and language model, our model is designed to characterize filler phrases depending on the speaking-style, thus can be trained with large corpora of different topics but the same style. Mutual information criterion is used to select topic independent filler words and their N-gram model is used for verification of key-phrase hypotheses. A dialogue-style dependent filler model improves the key-phrase detection in different dialogue applications. A lecture-style dependent model is trained with transcriptions of various oral presentations by filtering out topic specific words. It performs much better verification of key-phrases uttered during lectures of different topics compared with the conventional syllable-based model and large vocabulary model.

1. INTRODUCTION

One of the vital issues in real-world applications of speech recognition is flexibility to handle unconstrained utterances. We have introduced a combined detection and verification framework[1] that focuses on identifying the semantically significant portions and rejects the out-of-task parts of input utterances. Utterance verification technique based on acoustic models is introduced to give confidence measures to hypothesized key-phrases. In this paper, we propose a topic independent lexical and language modeling to enhance the key-phrase detection and verification.

It is well-known that lexical and language models are effective for improving keyword detection and suppressing false alarms[2][3][4]. They have two roles of improving the alternate (=filler) model and providing statistical contexts for keywords[2].

Most of the conventional works use domain dependent lexical entries and language models that are trained with the same large corpus such as Switchboard or WSJ corpus. However, it is not a realistic assumption that sufficient data is available for every single task-domain in real-world applications. For example, in unconstrained conversations or in oral presentations, the topic will be different at every session. None the less, there are applications where designated key-phrases must be detected correctly.

Therefore, a topic independent approach for lexical and language modeling is studied. Instead of task-domain specific models, we only assume that the speaking-style is same, such as dialogue-style or lecture-style. Then, using corpora of multiple topics, we select lexical entries that are independent of topics and characteristic to the speaking-style based on an information-theoretic criterion. The resultant lexicon and language model can hopefully be applied to speech sessions of different topics as long as the speaking-style is maintained.

2. TOPIC INDEPENDENT MODELING OF FILLERS

2.1. Filler Model for Key-Phrase Verification

A filler model is intended to characterize typical patterns that accompany keywords or key-phrases. It has to satisfy the following requirements: (1) sufficient coverage, (2) small word perplexity, and (3) small complexity.

As for lexical coverage, it is shown that, once adequate coverage of the vocabulary is realized, varying the vocabulary size from medium (=several hundreds) to large (=thousands) does not affect the performance[3]. It suggests that a topic independent lexicon can suffice reasonable coverage toward data of any kinds of (including new) topics. Small complexity of the model is desirable to realize efficient recognition.

The same sort of models can be used for not only keyword spotting but also for utterance verification[5]. An output hypothesis of the recognizer (W) for an input X is tested, and accepted if its score $P(X|W)$ is better than that by the verification model $P(X|\lambda^V)$.

This is formulated as a likelihood ratio (LR) test.

$$LR = \frac{P(X|W)}{P(X|\lambda^V)} \simeq \log P(X|W) - \log P(X|\lambda^V)$$

The verification model λ^V is competitive to recognized candidates. It is defined by the filler model with the property mentioned above.

In many previous works, a general acoustic sink model or a phone/syllable network model is used to serve the purpose. Such a simple model is not sufficient to characterize non-key-phrase events better than key-phrases. It realizes complete coverage and high likelihood to both fillers and key-phrases, thus it is not suitable to discriminate them.

A large vocabulary statistical model tries to provide both adequate coverage and constraint on fillers. However, it includes so many redundant entries and is not efficient. Moreover, serious mis-match is possible when the topic of input speech is changed.

2.2. Topic Independent Modeling

Our goal is to provide a model that is sufficient for fillers and robust against variety of topics and vocabulary set.

Key property of the model is that it is constructed in a domain independent manner. Instead of a domain dependent lexicon and corpus, we assume the model is dependent on the speaking-style. People use similar phrases in making an information query dialogue whatever the content of the query is. And they use a different style in giving an oral presentation in public. Based on the assumption, we train the filler model with large corpora that are not domain specific as long as their tasks are similar and so are the speaking-styles.

For the purpose, we adopt an information-theoretic criterion, which is widely used for topic identification or topic extraction. Specifically, mutual information between a word w and topics T is computed. Suppose there are a set of topics $T = \{t_1, \ldots, t_n\}$, the mutual information $I(T; w)$ for a word w indicates non-uniformity of the frequency of the word w in various topics, or how much the word correlates with specific topics. Thus, it can be used to measure how significantly the word w can contribute to identify the topics.

$$\begin{aligned} I(T; w) &= H(T) - H(T/w) \\ &= \sum_T P(t_i) \log \frac{1}{P(t_i)} - \sum_T P(t_i/w) \log \frac{1}{P(t_i/w)} \end{aligned}$$

Unlike topic identification, we pick up the words that appear in various topics universally, that is whose $I(T; w)$ values are small. The resultant word set will give reasonable coverage to inputs of any topics and be robust against the change of the domain.

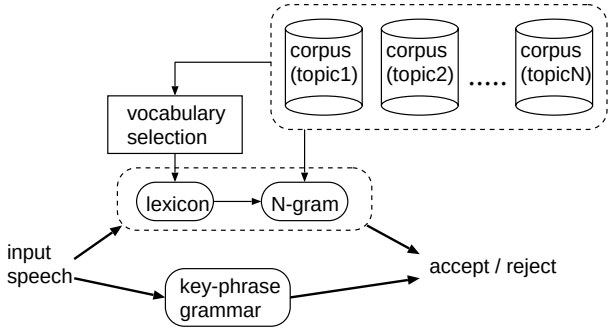

Figure 1: Process overview

For training procedure, we need corpora that consist of multiple topics. However, topic labels are not necessary since topic identification is not the purpose. We can simply use multiple corpora of different topics, or even split one corpus into segments and regard each segment to be of different topics. The first term of the above equation is to normalize the text size of each corpus or segment. The speaking-style of the corpora must be consistent, so that we can obtain an effective model unique to the style. Based on the generated lexicon, N-gram model is trained with the original corpora in order to incorporate more precise constraint.

The overview of the model training process and verification procedure is depicted in Figure 1. We present two applications of the modeling: dialogue-style model and lecture-style model.

3. KEY-PHRASE DETECTION IN DIALOGUE SPEECH

As a simple application and preliminary evaluation, we construct a filler model that is dependent on the dialogue-style. Specifically, we deal with information query which is a typical application of spoken dialogue systems. The purpose of the filler model is to improve the detection rate of key-phrases which will lead to robust speech understanding.

We made use of the ATIS-I corpus of 13099 utterances and Movie Locator corpus of 3777 utterances. The latter is a set of queries on movies being played and was collected at (former) AT&T Bell Labs. Since there are only two domains, the formulation in the previous section is almost equivalent to picking up common frequent words that appear in the both corpora. In this preliminary task, we use phrase modeling[6] that concatenates frequent word sequences, since estimating N-gram statistics for filler words is hard with this size

of data. As the result, 105 frequent filler phrases are selected.

The generated filler model is applied to speech understanding based on our key-phrase detection and verification approach[1]. It is used to generate competitive hypotheses in the detection process.

The evaluation was performed on 911 spontaneous utterances specifying locations (LOCATION sub-task) in a car reservation system, which were also collected at Bell Labs. Incorporation of the filler model improved the detection rate of key-phrases of locations from 40.1% to 58.4% for out-of-grammar samples that are not covered with the key-phrase grammar, while keeping the accuracy of 92.6% for in-grammar samples. The result demonstrates that our model trained with different corpora enhances the detection performance at a new task domain.

4. KEY-PHRASE VERIFICATION IN LECTURE SPEECH

Next, our modeling is applied to lecture-style speech. The task here is to detect and verify key-phrases uttered during oral presentations.

4.1. Filler Model Training

We made use of three corpora, all of which are transcriptions of panel discussions of different topics held at different occasions. The topics include spoken language processing, medical ethics, and local autonomy government. The text size of each corpus ranges from 10K to 30K words.

After removing those words that appear only once, the vocabulary size gets 2345. Then, the mutual information value $I(T; w)$ is computed to rank the words in ascending order.

Most of the functional words are ranked high. The typical verbs and adverbs in oral presentations are also included. Examples of nouns include *study, result, status, company, country, items, content, viewpoint, time, necessity*, though the texts are in Japanese. The domain specific words are clearly filtered out. The top 360 words are selected for the filler model lexicon. Then, word bigram model is trained using the corpora.

4.2. Evaluation on Utterance Verification

The filler model is applied to key-phrase verification for a slide projector operated with voice commands. Key-phrases are commands for the projector operation, such as "next slide" or "two slides back". They are represented as a finite state grammar. The vocabulary size for the commands is 56.

A lecturer uses the same microphone to give a presentation and to utter commands to the projector. Thus, most of input speech segments are not command key-phrases and contain vocabulary of over thousands.

A speech segment aligned with pauses is input to the recognizer that is made of subword HMM and the finite state grammar. It is also passed to the verifier of the subword-based filler model. The likelihood ratio (LR) of the two models are compared for final decision. If it is over a threshold, the input is accepted as a command. Otherwise it is regarded as a portion of lecture speech and discarded. The general subword HMM is trained with 20K sentence utterances by 132 speakers.

The speech samples for evaluation was collected from lectures given at our department. The topics and the speakers are different from those in the training material. The test set consists of 199 command key-phrase utterances and 646 speech segments of lectures whose duration lengths are comparable to those of key-phrases (less than 5 sec.).

For comparison of the verification model, we use a language model trained with Mainichi newspaper texts of 4 years (65M words), which is one of the largest Japanese corpora. The baseline method of free decoding with a syllable network is also tested.

The sum of the false acceptance (FA) rate and the false rejection (FR) rate is plotted against threshold values of the verification in Figure 2.

Use of a syllable network without lexical and language models gets the worst performance. However, the newspaper language model does not work effectively, either. The vocabulary of the newspaper is different from those in lectures, thus not suitable for discriminating lecture speech from commands. The filler model trained with various oral presentations significantly outperforms the above two methods. Even when we use the lexicon only and do not incorporate bigram statistics, the optimal performance is not lowered. Use of the bigram model makes the bottom of the operational curve flat, namely it realizes more robust verification against the thresholding condition.

Figure 3 plots the false rejection (FR) rate against the false acceptance (FA) rate. Since the false acceptance (FA) is more critical in this application, we should focus on the FR rate, where the FA rate is less than 1%. It also demonstrates that the proposed filler model is most effective. The bigram model is clearly better than the lexicon only in this graph. It realizes the FR rate of 3.0% at the FA rate of 0.2%, thus makes the voice-operated projector practical.

The improvement from the baseline syllable decoding is not less than the gain realized by the purely acoustical verification model, which was reported in [7]

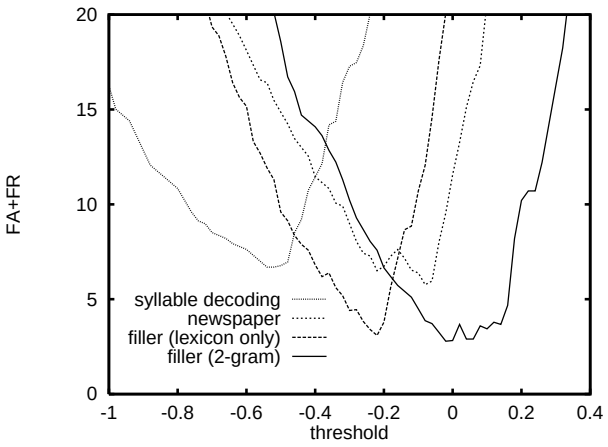

Figure 2: Comparison of verification models (FA+FR)

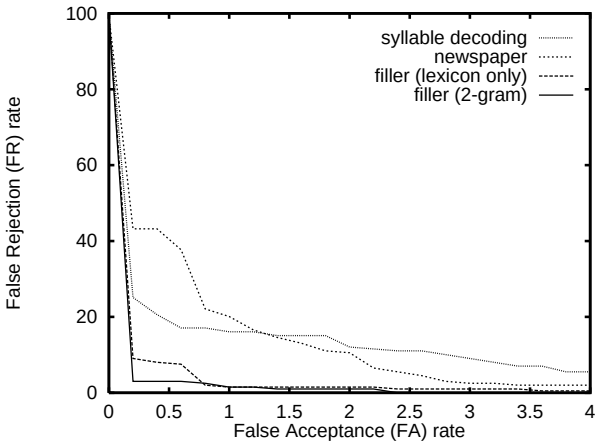

Figure 3: Comparison of verification models (FR vs. FA)

and adopted in our former work[1], although the task and database is not the same. It suggests possibility of combining the two approaches.

It is noteworthy that the verification model is made of very small complexity and works even more efficiently than the syllable decoding.

In Table 1, several variation of vocabulary sizes are compared. All of them use bigram statistics. The false rejection (FR) rates at the FA rate of 0.5% and 1.0% are figured out, since these points are critical as seen in Figure 3. The proposed vocabulary selection (MI: Mutual Information) method achieves the same performance as the original vocabulary of 2345 words, while reducing its size to 360. It is more effective than the simple vocabulary elimination method based on the word frequency. However, further reduction of the vocabulary increases verification errors.

Table 1: Comparison of vocabulary set

vocabulary size	False Rej. @FA=0.5%	False Rej. @FA=1.0%
2343 (original)	3.0	1.5
360 (MI)	3.0	1.5
100 (MI)	6.0	3.0
360 (frequent)	3.5	3.0

5. CONCLUSIONS

We have proposed the topic independent modeling of fillers for robust key-phrase detection and verification. It extracts a domain independent lexicon and N-gram statistics assuming the same speaking-style, thus can be trained with large corpora of different domains. Key property of the model is portability and generality. It is subword-based and can be ported to tasks of new topics without re-training. The model was realized in two different styles: dialogue-style and lecture-style. They were successfully applied to speech understanding and utterance verification, respectively.

Acknowledgment: The authors are grateful to Dr. Chin-Hui Lee at Bell Laboratories for his cooperation and comments to the work.

References

[1] T.Kawahara, C.-H.Lee, and B.-H.Juang. Flexible speech understanding based on combined key-phrase detection and verification. *IEEE Trans. Speech & Audio Process.*, 6(6):(to appear), 1998.

[2] J.R.Rohlicek, P.Jeanrenaud, K.Ng, H.Gish, B.Musicus, and M.Siu. Phonetic training and language modeling for word spotting. In *Proc. IEEE-ICASSP*, volume 2, pages 459–462, 1993.

[3] M.Weintraub. Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system. In *Proc. IEEE-ICASSP*, volume 2, pages 463–466, 1993.

[4] R.E.Meliani and D.O'Shaughnessy. Accurate keyword spotting using strictly lexical fillers. In *Proc. IEEE-ICASSP*, pages 907–910, 1997.

[5] R.C.Rose, B.-H.Juang, and C.-H.Lee. A training procedure for verifying string hypotheses in continuous speech recognition. In *Proc. IEEE-ICASSP*, pages 281–284, 1995.

[6] T.Kawahara, S.Doshita, and C.-H.Lee. Phrase language models for detection and verification-based speech understanding. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 49–56, 1997.

[7] R.A.Sukkar and C.-H.Lee. Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition. *IEEE Trans. Speech & Audio Process.*, 4(6):420–429, 1996.