Classifier-based Data Selection for Lightly-Supervised Training of Acoustic Model for Lecture Transcription

SHENG LI[1,a)] YUYA AKITA[2,b)] TATSUYA KAWAHARA[3,c)]

Abstract: The paper addresses a scheme of lightly-supervised training of acoustic model, which exploits a large amount of data with closed caption texts but not faithful transcripts. In the proposed scheme, a sequence of the closed caption text and the ASR hypothesis by the baseline system are aligned. Then, a dedicated classifier is designed and trained to select the correct one among them or reject both. It is demonstrated that the classifier can effectively filter the usable data for acoustic model training. The scheme realizes automatic training of the model with an increased amount of data. A significant improvement in the ASR accuracy is achieved from the baseline system and also in comparison with the conventional method of lightly-supervised training based on simple matching or confidence measure score.

1. Introduction

Automatic transcription of lectures is one of the promising applications of automatic speech recognition (ASR) as many courses of audio and video lectures are being digitally archived and broadcasted. Captions to the lectures are needed not only for hearing-impaired persons but also for non-native viewers and elderly people. ASR would also be useful for indexing the content.

ASR of lectures have been investigated for almost a decade in many institutions world-wide [1, 2, 3, 4, 5, 6, 7], but there are still technically challenging issues for the system to be practical level, including modeling of acoustic and pronunciation variations, speaker adaptation and topic adaptation. In this work, we address effective acoustic model training targeted on Chinese spoken lectures.

There is a large amount of audio and video data of lectures, but it is very costly to prepare accurate and faithful transcripts for spoken lectures, which are necessary for training acoustic and language models. We observed that, even given caption text, a lot of work is needed to make a faithful transcript because the caption text is much different from what is actually spoken and phenomena of spontaneous speech such as fillers and repairs need to be included.

In order to increase the training data for acoustic model, a scheme of lightly-supervised training, which does not require faithful transcripts but exploits available verbatim texts, has been explored for broadcast news [10, 11, 12] and parliamentary meetings [13]. In the case of TV programs, closed caption texts are used as a source for the scheme. A typical method consists of two steps. In the first step, a biased language model is constructed based on the closed caption text of the relevant program to guide the baseline ASR system to decode the audio content. The second step is to filter the reliable segments of the ASR output, usually by matching it against the closed caption or filtering with a threshold on the confidence measure score.

[1] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.
a) lisheng@ar.media.kyoto-u.ac.jp

The conventional filtering method, however, has a drawback that it significantly reduces the amount of usable training data. Moreover, it is presumed that the unmatched or less confident segments of the data are more useful than the matched segments because the baseline system failed to recognize them and may be improved with the additional training [12]. Recent work by Long et al. [14] proposed methods to improve the filtering by considering the phone error rate and confidence measures. Other researches, e.g. [15], introduced an improved alignment method for lightly-supervised training.

In this work, we propose to train a dedicated classifier to select the usable data for acoustic model training. Given an aligned sequence of the ASR hypothesis and closed caption text (and also reference text in the training phase), the classifier is trained based on a discriminative model to accept either the ASR result or the closed caption text, or reject both if they are not matched. It is trained with a database of a relatively small size used for training the baseline acoustic model and applied to a large-scale database that has closed caption texts but not faithful transcripts.

In the remainder of the paper, we first describe the corpus of Chinese spoken lectures and the baseline ASR performance in Section 2. Next, our proposed scheme of classifier design for lightly-supervised training is formulated in Section 3. Then, the implementation of the method to our lecture transcription task is explained and experimental results are presented in Section 4. The paper is concluded in Section 5.

2. Corpus and baseline ASR performance

For a comprehensive study on ASR of spontaneous Chinese language, we compile a corpus of Chinese spoken lectures and investigate the ASR technology using it.

2.1. Corpus of Chinese Lecture Room

While Chinese is one of the major languages for which ASR has been investigated, studies on Chinese lecture speech recognition are limited [8, 9], and a large-scale lecture corpus for this study has not been made. We have designed and constructed a corpus of Chinese spoken lectures based on the CCTV program of “Lecture Room” (百家講壇), which is a popular academic lecture program of China Central Television

(CCTV) Channel 10. Since 2001, a series of lectures have been given by prominent figures from a variety of areas.

As of the end of 2013, we made annotation (segmentation of the lecture part and faithful transcription) to the selected 98 lectures (90 speakers, 21 female, 69 male), which amount to 61.6 hours of speech and 1.2 M characters of text. They are categorized into three general topics: 38 lectures about history-culture-art, 29 lectures about society-economy-politics, and 31 lectures about science-technology. We have also collected 126 lectures with closed captions, which are not annotated so far. We call all of the data both annotated and unannotated as the Corpus of Chinese Lecture Room (CCLR). For the experimental purpose, we select 58 annotated lectures as the training set (CCLR-TRN), and 19 annotated lectures as the test set (CCLR-TST). The remaining 126 unannotated lectures are used for lightly-supervised training (CCLR-LSV). All these data sets are listed in Table 1.

Table.1 *Organization of CCLR corpus.*

	#lectures	Duration	Text Type
CCLR-TRN	58	35.2 hours	caption/faithful
CCLR-TST	19	11.9 hours	faithful
CCLR-LSV	126	62.0 hours	caption only

2.2. Baseline ASR system and performance

To build a baseline lecture transcription system, we used CCLR-TRN of 35.2 hours as the training set, and tested on CCLR-TST. We adopt 113 phonemes (consonants and 5-tone vowels) as the basic HMM unit. We use 39-dimensional PLP features with CMN+CVN for each speaker. The total number of tied triphone states is 3000 and each state has 16 Gaussian mixture components. Both MLE and MPE models are trained.

For the DNN model training, we use the same PLP features, and the only difference is the features are globally normalized to have a zero mean and a unit variance. We use the baseline MPE model to generate the state alignment label. The network has 429 nodes as input (5 frames on each side of the current frame), 3000 nodes as output and 6 hidden layers with 1024 nodes per layer. The training starts with unsupervised pretraining and followed by supervised fine-tuning based on frame-level cross-entropy training. When testing, the PLP features are feed-forwarded through the DNN network to generate the log-posterior probabilities, and then normalized by the state prior probabilities. The state prior probabilities are estimated from the training label. Julius 4.3 (DNN version) is used for decoding.

The dictionary consists of 53k lexical entries. The pronunciation entries are from the CEDICT open dictionary and the HKUST dictionary. The language model was built from faithful transcriptions of CCLR-TRN and other lecture texts collected from the web. We use our own decoder Julius 4.3.1 [16]. This baseline system achieved an average Character Error Rate (CER) of 39.31% with the MLE model, 36.66% with the MPE model and 31.60% with the DNN model for CCLR-TST.

3. Classifier design for data selection

3.1. Lightly supervised training framework

To perform the lightly supervised training, we need a criterion to select data. The conventional lightly supervised training relies on simple matching between the caption text and the ASR hypothesis, and thus discards so much data which could be useful. The other method is setting a threshold to the confidence measure score (CMS) of the utterances, but the CMS is not often reliable and tuning a threshold is not easy.

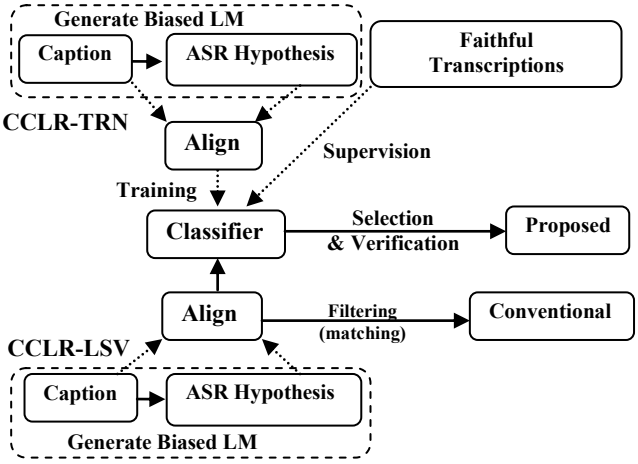

Fig.1 *Framework of proposed lightly supervised training.*

In this paper, we propose a data selection framework based on dedicated classifiers to replace the simple methods, as shown in Figure 1. Training of the classifiers is conducted by using the training database of the baseline acoustic model (CCLR-TRN). First, we generate the ASR hypothesis (1-best) using a biased language model which is made of the caption text interpolated with the baseline language model. Then, the ASR hypothesis is aligned with the corresponding caption text. By analyzing the aligned word sequence between the ASR hypothesis and the caption, we can categorize patterns by referring to the faithful transcriptions, as shown in Table 2.

Table.2 *Category of alignment patterns.*

	Caption		ASR Hypothesis		Faithful Transcriptions (reference)
C_1	发表	√	发表	√	发表
C_2	沦亡	X	沦亡	X	论文
C_3	雪山	X	学说	X	学术
C_4	雪辉	X	学会	√	学会
C_5	法人	√	发热	X	法人

(X means mismatching with reference, √ means matching)

- C1: the ASR hypothesis is matched with the caption and also a correct transcript. A majority of the samples falls in this category.
- C2: although the ASR hypothesis is matched with the caption, it is not a correct transcript. This case is rare.
- C3, C4, C5: the ASR hypothesis is different from the caption. In C3, neither of them are correct. In C4, the ASR hypothesis is correct. In C5, the caption is correct.

Note that the conventional method is equivalent to simply using C_1 and C_2. The objective of this study is to incorporate more effective data (C_4 and C_5) while removing erroneous data (C_2 and C_3).

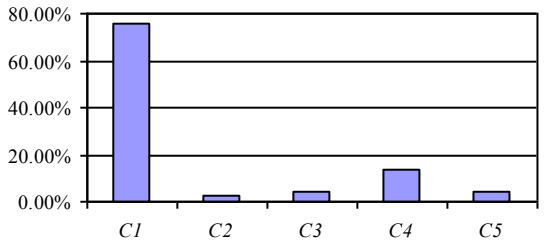

Fig.2 *Data distribution in CCLR-TRN.*

The distribution of these patterns in CCLR-TRN is shown in Figure 2. It is observed that 75.7% of them are categorized into *C1*. Among others, *C4* is the largest because the caption text is often edited from the faithful transcripts for readability. We initially tried to design a classifier to conduct classification of these five categories, but it turned to be difficult because of the complex decision and the data imbalance. Therefore, we adopt a cascaded approach.

3.2. Cascaded classifiers for word-level data selection

In the cascaded approach, we design two kinds of classifiers. One is for selection of the hypothesis and the other is for verification of the selected hypothesis.

C1 and *C2* are the matching cases between the ASR hypothesis and the caption. In these cases, the data selection problem is reduced to whether to accept or discard the word hypothesis. On the other hand, *C3, C4* and *C5* are the mismatching cases between the ASR hypothesis and the caption. We train a binary classifier to make a choice between the ASR hypothesis and the caption word. Then, we apply the other classifier to accept it. This classifier can be the same as the one used for *C1* and *C2*.

The classification is organized by the two binary classifiers in a cascaded structure as illustrated in Figure 3. Binary classifiers are focused on specific classification problems, so they are easily optimized. This design also mitigates the data imbalance problem. In Figure 3, one classifier is used for selection of the word hypothesis with highest credibility either from the ASR hypothesis or the closed caption, and the other is used for verification of the selected (or matched) hypothesis.

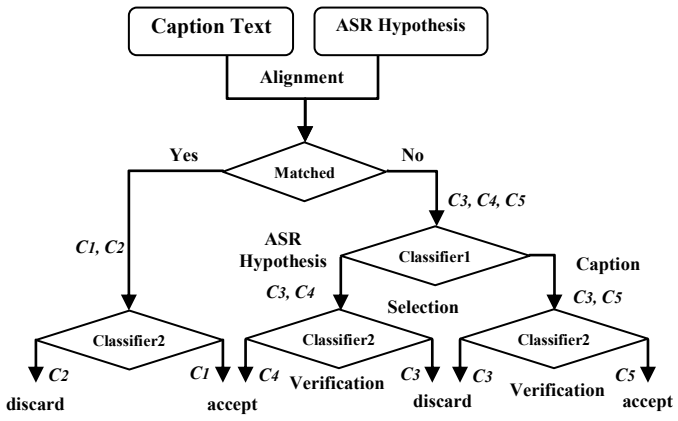

Fig.3 *Cascaded classifiers for data selection.*

3.3. Feature design for CRF

We use conditional random fields (CRF) [17] as the classifier for this task. It can model the relationship between the features and labels, considering sequential dependencies of contextual information. For this reason, it is used for many applications such as confidence measuring [18], ASR error detection [19], and automatic narrative retelling assessment [20].

When training the classifiers and conducting data selection, we need to convert the alignment patterns to a feature vector. A list of candidate features is shown in Table 3. These features include both acoustic and linguistic information sources. The text-based features are defined for both ASR hypothesis and caption text while the speech-based features are computed for the ASR result only.

For the target word W_0, both its ASR hypothesis and caption text have their own n-gram (NG), Part-of-Speech (POS) and tf-idf (TF) features. And these three groups of features are organized in pairwise forms. The other two groups of features, the confidence measure (CMS) and word duration (DUR), are shared by the ASR hypothesis and caption text of the target word W_0.

Moreover, The contextual information of the target word W_0 is also incorporated by adding features of the preceding two words (W_{-2} , W_{-1}) and following two words (W_1 , W_2). So the complete feature for a target word is over eighty dimensions.

Table.3 *Feature sets for classification.*

Feature Type	Features
Text-based	1. n-gram (NG), n=1,2,3
	2. Part-of-Speech (POS)
	3. tf-idf (TF)
Speech-based	1. confidence measure by decoder (CMS)
	2. word duration (DUR)

- The n-gram feature is a combination of word id and its log probability. And these features are organized for 1-gram, 2-gram and 3-gram.
- To get the Part-of-Speech (POS) feature, we trained a CRF classifier for POS tagging from Chinese-Tree-Bank (CTB) 4. We defined 15 Part-of-Speech tag symbols.
- The tf-idf (TF) feature is computed by multiplying the tf-value and the idf-value. The tf-value is calculated from the word frequency in the caption text of the current lecture. And the idf-value is computed by calculating how many documents include that word in the entire caption text inventory from CCLR-TRN and CCLR-LSV sets.
- The confidence measure (CMS) and word duration (DUR) feature are output by the baseline ASR system.

3.4. Utterance selection for acoustic model training

All the ASR hypotheses and the caption text are merged into a single word sequence after the matching and selection process, and every word in the sequence will have a label,

either "accept" or "discard", based on the verification process according to Figure 3.

Then, we need to make a decision whether or not this sequence of the data by the utterance unit is used for acoustic model training. Since acoustic model training is conducted based on phone labels rather than word labels, we compute a phone recall rate (PRR) for every utterance, which is the ratio of the number of the accepted phones to the total number of phones. In this process, errors in homonyms are tolerated.

It is not easy to figure out the optimal point between growth of noise and the amount of training data [22]. It is affected by a number of factors and often determined a posteriori depending on the data set and the baseline performance. In this work, we use only reliable utterances (PRR=100%) for lightly supervised training of acoustic model.

4. Experimental evaluations

4.1. Classifier implementation and performance

We first conduct speech segmentation to the utterance unit based on the BIC method [23] and speech clustering to remove non-speech segments and speech from other than the main lecturer in CCLR-LSV. The seed model comes from our baseline ASR system in Section 2. For each lecture, a biased language model is created by interpolating its closed-caption model and the baseline model with the weights 0.9 and 0.1.

In our experiment, we used CRF++ [24] to train two classifiers using CCLR-TRN: CRF-2, which is trained from {C_1, C_2}, and CRF-1, which is trained from {C_3, C_4, C_5}. In the implementation, we use second-order CRF. To make binary classification, we merge C_3 into C_4, because we observed the phone accuracy of ASR hypothesis is higher than that of the caption in C_3. Erroneous patterns in C_3 will be rejected by the second classifier CRF-2.

The performance of various feature combinations is compared by 5-fold cross validation on CCLR-TRN, as shown in Table 4. Performance measures are Precision, Recall and F-score:

$$\begin{cases} \text{Precision}=TP/(TP+FP) \\ \text{Recall}=TP/(TP+FN) \\ \text{F-score}=2\times\text{Precision}\times\text{Recall}/(\text{Precision}+\text{Recall}) \end{cases}$$

where TP is true positives, FP is false positives and FN is false negatives.

Table.4 *Feature set evaluation by 5-fold cross validation on CCLR-TRN.*

	CRF-1			CRF-2		
Feature	Precision	Recall	F-score	Precision	Recall	F-score
NG	0.73	0.75	0.74	0.99	0.99	0.99
POS	0.71	0.73	0.71	0.97	0.97	0.97
TF	0.67	0.71	0.69	0.96	0.96	0.96
NG+POS+TF	0.77	0.78	0.78	0.99	0.99	0.99
CMS	0.65	0.71	0.68	0.99	0.99	0.99
DUR	0.71	0.76	0.73	0.99	0.99	0.99
CMS+DUR	0.71	0.76	0.73	0.99	0.99	0.99
All Features	**0.79**	**0.80**	**0.79**	**0.99**	**0.99**	**0.99**

We get following conclusions from Table.4: Selection (CRF-1) is more difficult than verification (CRF-2). Feature combination is effective for CRF-1, while most of the features result in very high accuracy in CRF-2. CMS alone is not effective for CRF-1.

Therefore, we select the complete feature set. The features of five words (two preceding and two following) for both ASR hypothesis and caption text are concatenated to make a whole predictor vector. Although errors by CRF-1 in the first pass of the classification is inevitable, part of them are detected and discarded in the second pass of classification by using CRF-2.

4.2. ASR performance with enhanced model training

Next, we conduct lightly supervised training of the acoustic model by applying the classifiers to CCLR-LSV. Then, ASR performance is evaluated on CCLR-TST.

In this experiment, we use the same setting with the baseline system descripted in Section 2. We compare our proposed method with other three methods.

- "Baseline": the model trained by only using CCLR-TRN as mentioned in Section 2.
- "No-selection": simply pool the 58 CCLR-TRN lectures and 126 CCLR-LSV lectures together, and directly use the ASR hypothesis of CCLR-LSV without any selection.
- "Conventional": the conventional lightly supervised training which selects the data based on simple matching of the ASR hypothesis and the caption.

Table.5 *ASR performance (CER%) by lightly supervised training.*

	Durations (Hours)		Ave. CER% CCLR-TST		
	CCLR-TRN	CCLR-LSV	MLE	MPE	DNN
Baseline	35.2	0	39.31	36.66	31.60
No-selection	35.2	62.0	38.50	34.42	28.80
Conventional	35.2	26.5	38.51	34.68	29.19
Proposed	35.2	48.9	37.93	33.99	28.39

ASR performance in CER is listed for MLE models, MPE models and DNN models in Table 5. Experiment results show our proposed lightly supervised training method outperforms all other methods for MLE, MPE and DNN models. The improvement is statistically significant. The p-values from two-tailed t-test at 0.05 significant level of our proposed method compared with the Baseline and No-selection and Conventional methods 0.0031, 0.0017 and 0.028 for the MLE model, 1.96e-07, 0.011 and 3.28e-04 for the MPE model and 7.06e-09, 0.0183 and 0.0011 for the DNN model.

Another advantage of our method confirmed in this experiment is it can significantly enlarge the training data by selecting faithful data while discarding the mismatching segments effectively.

5. Conclusions

We have proposed a new data selection scheme for lightly supervised training of acoustic model. The method uses dedicated classifiers for data selection which are trained with the training database of the baseline acoustic model. We designed a cascaded classification scheme based on a set of binary classifiers, which are effectively trained with the relevant data set. Experimental evaluations demonstrate the

proposed lightly supervised training method effectively increase the faithful training data and improves the accuracy from the baseline model and in comparison with the conventional method.

References

[1] K.Maekawa, Corpus of Spontaneous Japanese: Its Design and Evaluation. In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 7-12, 2003.
[2] H.Nanjo and T.Kawahara. Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition. TSAP, Vol.12, No.4, pp.391-400, 2004.
[3] I.Trancoso, R.Nunes, L.Neves, C.Viana, H.Moniz, D.Caseiro, and A.I.Mata, Recognition of Classroom Lectures in European Portuguese. In Proc. INTERSPEECH, pp. 281-284, 2006.
[4] J.Glass, T.J.Hazen, S.Cyphers, I.Malioutov, D.Huynh, and R.Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In Proc. INTERSPEECH, pp. 2553-2556, 2007.
[5] H.Yamazaki, K.Iwano, K.Shinoda, S.Furui, and H.Yokota, Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition. In Proc. INTERSPEECH, pp. 2349-2352, 2007.
[6] T.Kawahara, Y.Nemoto, and Y.Akita, Automatic Lecture Transcription by Exploiting Slide Information for Language Model Adaptation. In Proc. ICASSP, pp.4929-4932, 2008.
[7] M.Paul, M.Federico, and S.Stucker, Overview of the IWSLT 2010 Evaluation Campaign. In Proc. IWSLT, pp. 3-27, 2010.
[8] J.Zhang, H.Chan, P.Fung and L.Cao. A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech. In Proc. INTERSPEECH, pp. 2781-2784, 2007.
[9] S.Kong, M.Wu, C.Lin, Y.Fu, and L.Lee. Learning on Demand - Course Lecture Distillation by Information Extraction and Semantic Structuring for Spoken Documents. In Proc. INTERSPEECH, pp. 4709-4712, 2009.
[10] L.Lamel, J.Gauvain, and G.Adda. Investigating Lightly Supervised Acoustic Model Training. In Proc. ICASSP, pp. 477-480, 2001.
[11] L.Nguyen and B.Xiang. Light Supervision in Acoustic Model Training. In Proc. ICASSP, Vol. 1, pp. I-185, 2004.
[12] H.Chan and P.Woodland. Improving Broadcast News Transcription by Lightly Supervised Discriminative Training. In Proc. ICASSP, Vol. 1, pp. 737-740, 2004.
[13] T.Kawahara, M.Mimura, and Y.Akita, Language Model Transformation Applied to Lightly Supervised Training of Acoustic Model for Congress Meetings. In Proc. ICASSP, pp.3853-3856, 2009.
[14] Y.Long, M.J.F.Gales, P.Lanchantin, X.Liu, M.S.Seigel and P.C.Woodland. Improving Lightly Supervised Training for Broadcast Transcription. In Proc. INTERSPEECH, 2013.
[15] J.Driesen and S.Renals. Lightly supervised automatic subtitling of weather forecasts. Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013.
[16] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In Proc. APSIPA ASC, pp.131-137, 2009.
[17] J.Lafferty, A.McCallum, and F.Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. ICML, 2001.
[18] M.S.Seigel and P.C.Woodland, Combining Information Sources for Confidence Estimation with CRF Models, In Proc. INTERSPEECH, 2011.
[19] W.Chen, S.Ananathakrishnan, R.Kumar, R.Prasad, and P.Natarajan, ASR error dection in a conversational spoken language translation system, In Proc. ICASSP, 2013.
[20] M.Lehr, I.Shafran, E.Prud'hommeaux, and B. Roark, Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment, In Proc. NAACL, 2013.
[21] H.Jiang, Confidence measures for speech recognition: a survey, Speech Communication, vol. 45, no. 4, pp. 455-470, Apr. 2005.
[22] H.Lin, and J.Bilmes, How to select a good training-data subset for transcription: submodular active selection for sequences, In Proc. INTERSPEECH, pp.2859-2862, 2009.
[23] M.Mimura, T.Kawahara, Fast Speaker Normalization and Adaptation Based on BIC for Meeting Speech Recognition, IEICE TRANSACTIONS on Information and Systems (Japanese Edition) vol.J95-D No.7.
[24] T. Kudo. CRF++ toolkit. http://crfpp.sourceforge.net/, 2005.