LANGUAGE MODEL TRANSFORMATION APPLIED TO LIGHTLY SUPERVISED TRAINING OF ACOUSTIC MODEL FOR CONGRESS MEETINGS

Tatsuya Kawahara   Masato Mimura   Yuya Akita

Kyoto University, Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

For effective training of acoustic and language models for spontaneous speech such as meetings, it is significant to exploit the texts available in a large scale, which may not be faithful transcripts of the utterances. We have proposed a language model transformation scheme to cope with the differences between verbatim transcripts of spontaneous utterances and human-made transcripts such as those in proceedings. In this paper, we investigate its application to lightly supervised training of the acoustic model. By transforming the corresponding text in the proceedings, we can generate a very constrained model to predict the actual utterances. The experimental evaluation with the transcription system for the Japanese Congress meetings demonstrated that the proposed scheme can generate accurate labels for acoustic model training and thus realizes the comparable ASR (Automatic Speech Recognition) performance to the case using manual transcripts.

***Index Terms*—** speech recognition, language model, acoustic model, lightly supervised training

1. INTRODUCTION

Recently, the major target of LVCSR (Large-Vocabulary Continuous Speech Recognition) systems has been shifting to spontaneous speech such as conversations, lectures and meetings. One of the most fundamental problems in training an acoustic model for this kind of spontaneous speech is the insufficient amount of training data to cover wide variation of the acoustic features. It is very difficult and costly to prepare a large speech corpus, because it involves manual transcription of utterances with many disfluencies, compared to the reading of prepared materials.

To address this problem, the approach of lightly supervised training has been proposed and investigated[1]. The approach exploits texts, which are not faithful transcripts of what was spoken, but “cleaned” by human transcribers, for generating labels for acoustic model training. Most of the conventional works made use of closed caption texts in broadcasts in two ways; first, to generate a biased language model to automatically transcribe the speech, and then to filter the recognition hypotheses by aligning with the caption texts. On the other hand, Chan et al.[2] reported that the second step of the data selection was not useful, and all the data should be used instead. And Nguyen et al.[3] pointed out that it is vital to obtain the words the recognizer could have failed when using a baseline model.

By considering these findings, in this paper, we address effective label generation for lightly supervised training. We have been developing an automatic transcription system for the Japanese Congress (Diet) meetings[4]. Unlike European Parliament Plenary Sessions (EPPS), which were targeted by the TC-Star Project[5][6][7], the great majority of the sessions in the Japanese Congress are in committee meetings. They are more interactive and thus spontaneous, compared to plenary sessions. This means that there are many differences between the actual utterances and the final text in the proceedings or minutes. In our preliminary analysis, it is observed that an edit distance between the two is around 11%. Lightly supervised training is explored in the EPPS recently by Paulik et al.[8], but they simply use the final text edition (FTE) as it is, although they try to use the translated version as well.

We present a method which estimates the language model to accurately predict the actual utterances from the proceedings text. We have proposed a scheme of language model transformation and demonstrated its effectiveness in the baseline language modeling[4]. In this paper, the scheme is applied to the transcript generation for lightly supervised training of the acoustic model, so that we do not have to prepare manual transcripts for updating the model.

The remainder of this paper is organized as follows. Section 2 gives an overview of the proposed scheme of language model transformation. In Section 3, we describe lightly supervised training of the acoustic model based on the language model transformation. Experimental evaluations with the transcription system for the Japanese Congress meetings are presented in Section 4. Finally, Section 5 concludes the paper.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

2. STATISTICAL TRANSFORMATION OF LANGUAGE MODEL

We have proposed a scheme of language model transformation to cope with the differences between spontaneous utterances (verbatim text: V) and human-made transcripts (written-style text: W)[9]. In this scheme, the two are regarded as different languages and statistical machine translation (SMT) is applied. It can be applied in both directions: to convert a faithful transcript of the spoken utterances to a document-style text, and to recover the faithful transcripts from the human-made texts.

The decoding process is formulated in the same manner as SMT, which is based on the following Bayes' rule.

$$p(W|V) = \frac{p(W) \cdot p(V|W)}{p(V)} \qquad (1)$$

$$p(V|W) = \frac{p(V) \cdot p(W|V)}{p(W)} \qquad (2)$$

Here the denominator is usually ignored in the decoding.

Note that, in this case, the process to uniquely determine V (eq. 2) is much more difficult than the cleaning process (eq. 1) because there are more arbitrary choices in this direction; for example, fillers can be randomly inserted in (eq. 2) while all fillers are removed in (eq. 1). Therefore, we are more interested in estimating the statistical language model of V, rather than recovering the text of V. Thus, we derive the following estimation formula.

$$p(V) = p(W) \cdot \frac{p(V|W)}{p(W|V)} \qquad (3)$$

The key point of this scheme is that the available text size of the document-style texts W is much larger than that of the verbatim texts V needed for training ASR systems. For the Congress meetings, we have huge archives of proceedings texts. Therefore, we fully exploit their statistics $p(W)$ to estimate the language model $p(V)$ for ASR.

The transformation is actually performed on occurrence counts of N-grams as below.

$$N_{gram}(v_1^n) = N_{gram}(w_1^n) \cdot \frac{p(v|w)}{p(w|v)} \qquad (4)$$

Here v and w are individual transformation patterns. We model substitution $w \rightarrow v$, deletion of w, and insertion of v, by considering their contextual words. [1] $N_{gram}(w_1^n)$ is an N-gram entry (count) including them, thus to be revised to $N_{gram}(v_1^n)$. Estimation of the conditional probabilities $p(v|w)$ and $p(w|v)$ requires an aligned corpus of verbatim transcripts and their corresponding document-style texts. We have constructed the "parallel" corpus by using a part of the proceedings of the Japanese Congress meetings. The conditional probabilities are estimated by counting the corresponding patterns observed in the corpus. Their neighboring words are taken into account in defining the transformation patterns for precise modeling. For example, an insertion of a filler "ah" is modeled by $\{w = (w_{-1}, w_{+1}) \rightarrow v = (w_{-1}, ah, w_{+1})\}$, and the N-gram entries affected by this insertion are revised. A smoothing technique based on POS (Part-Of-Speech) information is introduced to mitigate the data sparseness problem. Please refer to [4] for implementation details and evaluation. The approach was shown to be effective in language modeling for the automatic meeting transcription system.

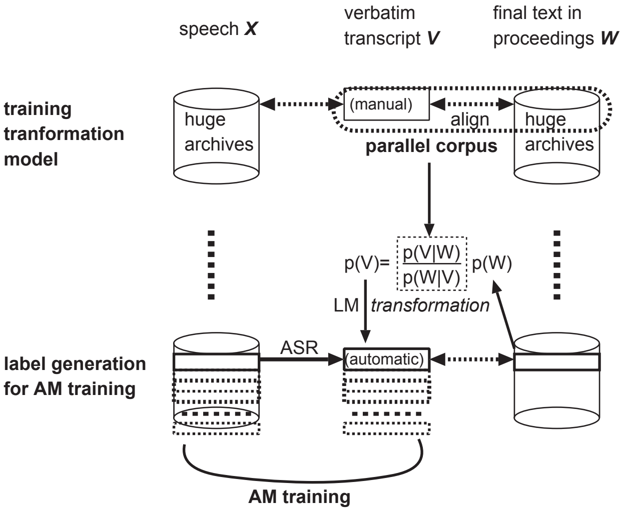

Fig. 1. Procedure overview of the proposed method

3. LIGHTLY SUPERVISED TRAINING OF ACOUSTIC MODEL

In this paper, we apply the language transformation scheme to lightly supervised training of the acoustic model. For the Congress meetings, we have large archives of speech which are not faithfully transcribed but have edited texts in the proceedings. Since it is not possible to uniquely recover the original verbatim transcript from the proceedings text, as mentioned in the previous section, we generate a dedicated language model for decoding the speech using the corresponding proceedings text. As a result of ASR, we expect to obtain a verbatim transcript with high accuracy.

The whole process is depicted in Fig. 1. For each turn (typically ten seconds to three minutes, and on the average one minute) of the meetings, we compute N-gram counts from the corresponding final text in the proceedings. Here, we adopt a turn as a processing unit, because the whole session (typically two to five hours) is too long, containing a variety of topics and speakers. A preliminary experimental

[1] Unlike ordinary SMT, permutation of words is not considered in this transformation.

comparison is presented in the next section. Automatic topic segmentation[10] may be incorporated for precise modeling, but we did not use it in this work.

The N-gram entries and counts are then converted to the verbatim style using the transformation model (eq. 4). Here we count up to trigrams. Insertion of fillers and omission of particles in the N-gram chain are also modeled considering their context in this process.

Then, ASR is conducted using the dedicated model to produce a verbatim transcript. The model is very constrained and still expected to predict spontaneous phenomena such as filler insertion. For acoustic model training, phone sequences are required. To cope with pronunciation variation in spontaneous speech, possible pronunciation variation patterns (surface forms) are predicted based on our generalized statistical model[11], and they are added to the dictionary used in ASR.

The best phone hypothesis is used as the label for the standard HMM training based on ML (Maximum Likelihood) criterion. For discriminative training such as MPE (Minimum Phone Error) criterion, we also generate competing hypotheses using a baseline language model. It is also possible to explore further correction of the ASR result by referring to the cleaned text[12].

4. EXPERIMENTAL EVALUATION

The proposed scheme has been implemented and evaluated with the automatic transcription system for the Japanese Congress meetings. The training corpus for the language transformation model and the baseline acoustic model via the standard supervised training was built from dozens of meetings in the years 2003-2005. It consists of 134 hour speech and it was faithfully transcribed and aligned with the final text in the proceedings to train the transformation model. The text size is about 1.8M words.

The acoustic features were 12-dimensional MFCCs and a power together with their Δ and $\Delta\Delta$. CMN (Cepstral Mean Normalization), CVN (Cepstral Variance Normalization), and VTLN (Vocal Tract Length Normalization) were performed. The acoustic model is triphone HMMs, which consist of 5000 shared states, each having 32 Gaussian components. We also incorporate the MPE training.

4.1. Evaluation on Label Generation

The additional training data for lightly supervised training of the acoustic model were prepared from the meetings in the years of 2006 and 2007. For these years, the general election was held and then the prime minister and cabinet members were replaced, thus many of major speakers in the meetings were different from those of the training corpus. In this experiment, we aim to update the acoustic model using these speech data. The data consist of 26 sessions and 91 hour speech in total.

Table 1. Accuracy of transcripts for acoustic model training

unit	method	Corr.	Acc.
	baseline	82.3%	79.5%
session	proceedings	83.6%	81.3%
	proc. + baseline	86.5%	83.9%
	proposed method	85.6%	83.3%
turn	proceedings	86.1%	83.5%
	proc. + filler	88.7%	86.2%
	proposed method	91.6%	89.9%

The proposed method was applied to generate transcripts for them. We first investigate the accuracy of the transcripts, which is measured by the word accuracy. Faithful transcripts were manually prepared for this evaluation. We made a comparison of the turn-based modeling with the case using a single model for the whole session. For the whole session, we can generate a biased language mode by interpolating the final text of proceedings with the baseline model. This is the conventional method widely used in the previous works. Note that applying this method to the turn-based modeling is not practical because there are a huge number of turns and the size of the biased language model is very large (1.6GB in this experiment). On the other hand, the proposed language model transformation method generates a very compact model (100MB), thus can be easily applied to many turn units.

For comparison within the turn-based modeling, we generated two different language models dedicated to the label generation. One model (proceedings) simply used the corresponding final text of proceedings. This corresponds to the conventional method such as [8]. The other (proc + filler) incorporated filler entries into the language model. Specifically, we added all filler entries to the lexicon, and used the unigram statistics of the training corpus by discounting the pause entry in the proceedings model. This is a simple version of the proposed method by focusing on the fillers and disregarding their context.

Table 1 lists the word percent correct (Corr.) and accuracy (Acc.). It is apparent that the turn-based models result in significantly higher accuracy than the whole-session models, because they provide stronger constraints. With the proposed language transformation method, we could recall 92% of the words. The accuracy is much higher than the case simply using the proceedings by 6% absolute. It is also confirmed that the proposed statistical model outperforms the simple filler-insertion model.

4.2. Evaluation on ASR

Then, we investigate the effect of the lightly supervised training on ASR using another test set of two sessions of the budget committee, which were held in February 2008. They con-

Table 2. ASR performance for test-set (WER)

session	ML training	MPE training
baseline model	14.6%	13.2%
proposed method	14.2%	12.6%
reference (manual)	14.1%	12.4%

stitute 2.5 hour speech of 121 turns. The language model for ASR was trained based on the transformation from large archives of proceedings whose text size is 87M words. The vocabulary size is 52K. Our decoder Julius[2] was used for LVCSR.

Table 2 shows the recognition performance by the Word Error Rate (WER) in the cases of ML training and MPE training. The baseline is the result of the baseline model without incorporating the additional data for acoustic model training. The proposed method of lightly supervised training reduced the WER by 0.6% absolute in the MPE training, which is a statistically significant improvement. For reference, the acoustic model was trained with the same data using the manual transcripts, and the recognition performance was much the same as that of the proposed method. This result demonstrates that we can update the acoustic model without manual transcription. [3]

In this evaluation setup, the size of the additional data was smaller than that of the baseline training corpus, since the primary goal was updating the model. We expect that the effect by the proposed method will be larger when a much larger amount of data are provided in a longer period.

5. CONCLUSIONS

We have presented a language model transformation scheme applied to lightly supervised training of the acoustic model. In the experimental evaluation, it is confirmed that the proposed method can generate accurate labels for the model training, and thus the acoustic model can be updated effectively, realizing the comparable performance to the case using the manual transcripts. The technique is vital for updating the acoustic model by reflecting the change in the group of members of the Congress, together with updating the language model to reflect new topics, which can also be automatically done with the transformation model.

We will investigate the effect of the scheme in a long term when the ASR module is deployed in the Japanese Congress. We also plan to apply the proposed scheme to the lecture transcription system.

Acknowledgment: The work is partially supported by Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communications, Japan.

[2] `http://julius.sourceforge.jp/`

[3] We could not conduct evaluation of the other methods in Table 1 in lightly supervised training since it would take too much time.

6. REFERENCES

[1] L.Lamel, J.Gauvain, and G.Adda. Investigating lightly supervised acoustic model training. In *Proc. IEEE-ICASSP*, volume 1, pages 477–480, 2001.

[2] H.Y.Chan and P.Woodland. Improving broadcast news transcription by lightly supervised discriminative training. In *Proc. IEEE-ICASSP*, volume 1, pages 737–740, 2004.

[3] L.Nguyen and B.Xiang. Light supervision in acoustic model training. In *Proc. IEEE-ICASSP*, volume 1, pages 185–188, 2004.

[4] Y.Akita and T.Kawahara. Topic-independent speaking-style transformation of language model for spontaneous speech recognition. In *Proc. IEEE-ICASSP*, volume 4, pages 33–36, 2007.

[5] C.Gollan, M.Bisani, S.Kanthak, R.Schluter, and H.Ney. Cross domain automatic transcription on the TC-STAR EPPS corpus. In *Proc. IEEE-ICASSP*, volume 1, pages 825–828, 2005.

[6] J.Loeoef, M.Bisani, C.Gollan, G.Heigold, B.Hoffmeister, C.Plahl, R.Schlueter, and H.Ney. The 2006 RWTH parliamentary speeches transcription system. In *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, pages 133–138, 2006.

[7] B.Ramabhadran, O.Siohan, L.Mangu, G.Zweig, M.Westphal, H.Schulz, and A.Soneiro. The IBM 2006 speech transcription system for European parliamentary speeches. In *Proc. INTERSPEECH*, pages 1225–1228, 2006.

[8] M.Paulik and A.Waibel. Lightly supervised acoustic model training EPPS recordings. In *Proc. INTERSPEECH*, pages 224–227, 2008.

[9] Y.Akita and T.Kawahara. Efficient estimation of language model statistics of spontaneous speech via statistical transformation model. In *Proc. IEEE-ICASSP*, volume 1, pages 1049–1052, 2006.

[10] Y.Akita, Y.Nemoto, and T.Kawahara. PLSA-based topic detection in meetings for adaptation of lexicon and language model. In *Proc. INTERSPEECH*, pages 602–605, 2007.

[11] Y.Akita and T.Kawahara. Generalized statistical modeling of pronunciation variations using variable-length phone context. In *Proc. IEEE-ICASSP*, volume 1, pages 689–692, 2005.

[12] S.Petrik and G.Kubin. Reconstructing medical dictations from automatically recognized and non-literal transcripts with phonetic similarity matching. In *Proc. IEEE-ICASSP*, volume 4, pages 1125–1128, 2007.