LIGHT SUPERVISION IN ACOUSTIC MODEL TRAINING

Long Nguyen, Bing Xiang

BBN Technologies
10 Moulton St., Cambridge, MA 02138, USA
{ln,bxiang}@bbn.com

ABSTRACT

In this paper, we present a new *light supervision* method to automatically derive additional acoustic training data for broadcast news transcription systems. In this method, a subset of the TDT corpus, which consists of broadcast audio with corresponding closed-caption (CC) transcripts, is identified by aligning the CC transcripts and the hypotheses generated by lightly-supervised decoding. Phrases of three or more contiguous words, that both the CC transcripts and the decoder's hypotheses agree, are selected. The selection yields 702 hours, or 72% of the captioned data. When adding 700 hours of selected data to the baseline 141-hour broadcast news training data set, we achieved a 13% relative word error rate reduction. The key to the effectiveness of this light supervision method is the use of a *biased* language model (LM) in the lightly supervised decoding. The biased LM, in which the CC transcripts are added with a heavy weight, helps in selecting words the recognizer could have misrecognized if using a fair LM.

1. INTRODUCTION

Statistical modeling is the dominant approach in state-of-the-art large vocabulary speech recognition systems. To have a reliable and accurate estimation of the speech models, large amount of manually transcribed training data is usually expected. However, it is expensive to have accurate orthographic transcripts for large speech corpora because manual transcription is a time-consuming task. To tackle this problem, lightly supervised acoustic model training on data having less accurate transcripts was proposed in [1]. In this study, the automatic transcripts generated by a recognition system were filtered with the CC transcripts so that only a subset of the data was chosen. A slight improvement in recognition accuracy was obtained compared to the approach of simply training on all available data with automatic transcripts. The largest training set reported in that work is around 200 hours. Recently, experiments with acoustic models trained on the 1400-hour TDT (Topic Detection and Tracking) corpus were reported to show substantial recognition error reduction [2].

At BBN, we have been experimenting with a different light supervision method in acoustic model training using the TDT corpus. Even though our method follows the same high-level procedure – decode then filter – used in [1] and [2], it seems that the strength of our method is drawn from these two facts: (1) the use of a *biased* LM in the decoding step, and (2) the strict criterion of selecting only phrases of three or more contiguous words such that both the CC transcripts and the decoder's hypotheses agree. The biased LM, in which the CC transcripts are added with a heavy weight, helps in selecting words the recognizer could have misrecognized if using a fair LM.

The paper is organized as follows. In Section 2, we describe the TDT corpus. The data selection procedure is introduced in Section 3. The brief description of our BN transcription system is presented in Section 4. We then report experimental results in Section 5. We conclude in Section 6.

2. TDT CORPUS

The TDT corpus was set up by LDC for research on Topic Detection and Tracking. The main portion of the corpus consists of audio data with corresponding CC transcripts. The corpus contains three subsets: TDT2, TDT3, and TDT4; each collected in a different time period as shown in Table 1. The TDT2 subset contains data from four sources: *ABC World News Tonight, CNN Headline News, PRI The World,* and *VOA English News*. Two more sources, *MSNBC News with Brian Williams* and *NBC Nightly News*, were added in the TDT3 and TDT4 subsets. They are somewhat 'representatives' of the main broadcasting media in the US: radio (PRI, VOA), broadcast TV (ABC, NBC), and cable TV (CNN, MSNBC). The total amount is about 1400 hours.

Subset	Period	Sources	Shows	Hours
TDT2	01/1998-06/1998	4	1034	633
TDT3	10/1998-12/1998	6	731	475
TDT4	10/2000-01/2001	6	425	294

Table 1: Statistics of the three subsets of the TDT corpus

Closed-caption transcripts of broadcast news programs are known to contain errors when compared to verbatim transcripts. One of the serious errors which prohibit the direct use of CC transcripts in acoustic model training for automatic speech recognition is the omission or paraphrasing of words. For example, a passage of the CC transcript of an ABC World News Tonight show reads *"The Republican leadership council is going to air ads promoting Ralph Nader"*. However, the announcer actually said *"The Republican leadership council,* **a moderate group,** *is going to air ads promoting* **the Green Party candidate** *Ralph Nader"*. It is very likely that, if the CC transcript of this sentence is used for acoustic model training, it would be rejected due to the failure in aligning the text to the audio or it would corrupt the phonetic models by being forcefully aligned to the wrong phonemes.

3. SELECTION PROCEDURE

The selection procedure is depicted in Figure 1. The CC transcripts are normalized first to create the reference transcripts in STM and SNOR formats. The SNOR transcripts are then added with a heavy weight to the original Hub4 LM to generate a biased LM to be used in the lightly-supervised decoding step. The TDT audio data are then decoded using the biased LM and the original Hub4 acoustic

model (AM). The decoder's hypotheses are then scored against the CC reference transcripts in STM format using the standard scoring script *sclite* provided by NIST. The alignments of the decoder's hypotheses and the CC transcripts during scoring, word by word with time-stamps, are captured in one of the *sclite's* outputs in SGML format. The selection step is done by searching through the alignment and identifying the completely matched phrases to be selected as additional acoustic training data.

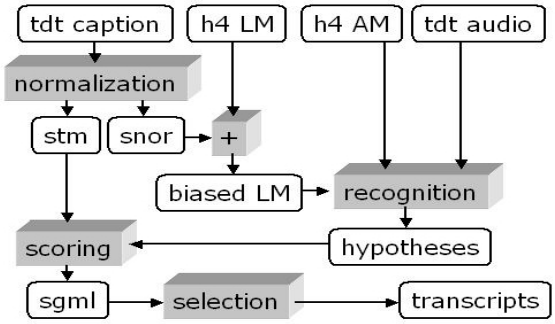

Figure 1: Diagram of the automatic selection procedure

3.1. Data Pre-Processing

The CC transcript, in SGML format, of a TDT show is organized chronologically as a sequence of segments corresponding to topics or stories. Each segment is tagged (or time-stamped) with time offsets when it begins and when it ends. These SGML tags are used to split the transcript into turns and sentences. The written text is normalized into SNOR format to be used in building the biased LM for lightly-supervised decoding later.

The SGML-format CC transcript is also transformed into STM format to be used as *truth reference* for scoring of the decoder's hypotheses during the selection step. Note that we exclude segments tagged as "MISCELLANEOUS" in the original CC transcripts, assuming they are probably commercials or transitional *chit-chat* speech among announcers. We are now unsure if that is a wise decision.

3.2. Biased Language Models

We define a *biased* LM as one in which the truth reference of the audio we plan to decode is used with some favorable weight in training that LM. There are many possibilities in constructing such biased language models: show-, subset-, or corpus-specific. Another research issue is how much data from sources other than the TDT transcripts should be used in building the language models.

We reiterate that the CC transcripts contain errors. A good recognizer might hypothesize correct words at those erroneous regions but it is not simple to automatically confirm that it is true. However, that same recognizer might make mistakes at other places. After all, the word error rate (WER) of state-of-the-art BN transcription systems is 10+%. Lower WERs can be achieved by using a LM with a certain bias towards the known-in-advance CC transcripts. However, if the bias is too strong, the recognizer might repeat the same errors existing in the CC transcripts. So, the art is in how to balance the bias in such a way that the recognizer can confirm the correct words and, at the same time, point out the errors in the CC transcripts. Recall that we use lightly-supervised decoding in order to generate a *second opinion* whether a word in the CC transcripts is correct or not. If the recognizer hypothesizes the same word as in the CC transcript, it is likely that that word is correct. Otherwise, there is a chance that the word is an error and should not be used as training data.

Three biased language models were built for the three subsets of the TDT corpus. A portion of the English GigaWord news text corpus published by LDC was used as the *fair* source of LM training data. Specifically, we used the 1998-2000 data from the New York Times Newswire Service, the Associated Press World-stream English Service, and about four months of 1998 data from the Los Angeles Times and the Washington Post newspapers. In total, this fair source consists of about 360M words. A common trigram LM was trained using this 360M-word fair source. A subset-specific trigram LM was estimated for each subset by merging (or interpolating) the subset's transcripts with a weight of 4 with the common LM. The lexicon used in these LMs has about 40K words derived from the most frequent words of the fair source. New words from the TDT corpus were also added if they have phonetic pronunciations in our 80K-word master dictionary.

3.3. Lightly-Supervised Decoding

Each subset of the TDT corpus was decoded using a BN recognizer similar to the one described in [3], but with a subset-specific biased LM. The band-specific and gender-dependent acoustic models were trained on the Hub4 141-hour training data set. Each show was decoded separately as if it was a new test set. The overall word correct rate of each show, measured against the CC transcript, ranged from 80% to 90%. [WER is not a good measurement in this work because of the high insertion rate at commercial segments that have no captions.] The decoding time was about 10x real time. [If the decoding of all 1400 hours of the TDT corpus was run on a single processor, it would take about a year and a half! (1400*10/24/365 = 1.59)]

3.4. Selection

We used NIST's *sclite* scoring tool to align the decoder's hypotheses against the CC reference transcripts. The only reason to use *sclite* is to take advantage of its ability to output the time-stamped word alignments, word by word in SGML format with a tag of either being correct or incorrect (i.e. substitution, insertion, or deletion). The selection process was carried out by searching through the SGML file to identify phrases of three or more contiguous words that both the CC transcript and the decoder's hypothesis agree. Utterances of one/two words were also selected if there were no errors. As shown in Table 2, totally 702 hours of data were selected, which is 72% of the captioned data and 50% of the raw data. We think that the selection yield rate is reasonable.

Subset	Raw	Captioned	Selected
TDT2	633	425	305
TDT3	475	328	241
TDT4	294	213	156
All	1402	966	702

Table 2: Selection results from TDT data (in hours)

The main reason to use the criterion of three or more contiguous words being matched is based on the fact that all trigrams occuring in the CC transcripts exist in the biased LM. The *second opinion* from the decoder after *listening* to the audio, albeit being biased towards expecting these trigrams, confirms that it is likely the phrase in the CC transcript is correct. Selecting phrases of two or one matched words while the neighboring words on the left and right of the phrase being unmatched seems to be risky.

4. THE BBN BN TRANSCRIPTION SYSTEM

At the core of the BBN BN Transcription system is the Byblos multi-pass recognizer. Various acoustic and language models at different levels of sophistication are deployed at different passes and/or stages. Even though this work is about how to improve the acoustic models using additional data derived automatically through light supervision, we briefly describe the BBN BN transcription system here to provide sufficient context.

4.1. Recognizer

The Byblos multi-pass recognizer [4] first does a fast match of the data to produce scores for numerous word endings using a coarse phonetically-tied-mixture (PTM) AM and a bigram LM. Next, a state-clustered tied-mixture (SCTM) AM and an approximate trigram LM are used to generate N-best hypotheses. N-best hypotheses are then re-scored and re-ranked using a cross-word SCTM AM and a 4-gram LM. The top-1 of the re-ranked N-best hypotheses is the recognition result. In other words, the decoding process is a three-step sequence (fast-match, N-best generation, and N-best rescoring) with finer-detailed models being used on narrower search space at later steps [5].

The decoding process is repeated three times. First, speaker-independent (and gender-independent) acoustic models are used in the decoding to generate hypotheses for unsupervised adaptation. Then, the decoding is repeated but with speaker-adaptively-trained acoustic models that have been adapted to the hypotheses generated in the first stage. The last decoding is similar to the second but acoustic models are adapted to the second stage's hypotheses using larger numbers of regression classes.

4.2. Acoustic Model Training

The acoustic model training procedure used at BBN can be logically described in these four stages.

Front-end Processing: 14-dimensional Perceptual Linear Predictive (PLP) [6] cepstral coefficients are extracted from the overlapping frames of audio data with a frame rate of 10ms. Cepstral mean subtraction is applied for normalization. The normalized energy is used as the 15^{th} component. In addition, the first, second, and third derivatives of the 15 components are also used to form a 60-dimensional feature vector.

ML-SI Training: The 60-dimensional feature vectors are transformed into 46-dimensional vectors by using a global HLDA and diagonalizing transform. The speaker-independent AMs (i.e. PTM, SCTM, and cross-word SCTM) are trained using the EM algorithm. These models are to be used in the SI decoding stage.

ML-HLDA-SAT: Speaker-dependent HLDA transforms [7] are then estimated in the original 60-dimensional space to project the feature vectors into another 46-dimensional feature space. The reduced feature space is further refined by using Constrained Maximum Likelihood Linear Regression (CMLLR) adaptation [8]. The speaker-adaptively-trained AMs are then trained using the maximum-likelihood (ML) criterion. These models are subsequently referred to as HLDA-SAT models and to be used only in the adapted decoding stages.

MMI-HLDA-SAT: In the last stage of acoustic model training, all training data is decoded using the HLDA-SAT models to generate lattices. Then a new set of AMs are estimated using these lattices under the MMI criterion [9]. These models are subsequently referred to as MMI models and can be used in place of the HLDA-SAT models. Note that this last stage of training is optional within Byblos since it is computationally expensive.

4.3. Language Models

The lexicon used in this system contains 61K words with about 4% of the words having more than one pronunciation. The out-of-vocabulary rate measured on the development test set is 0.35%. A total of 1.3 billion words selected from various LM text corpora released by LDC or collected at BBN are used to train the trigram and 4-gram language models. The trigram LM used in the recognizer for N-best generation has about 42M trigrams and 19M bigrams. The 4-gram LM used in N-best rescoring has about 730M 4-grams and the perplexity on the BN development test set is 138. This is the same set of language models used in the BBN EARS RT03 Evaluation system [10].

5. EXPERIMENTAL RESULTS

We carried out a sequence of experiments to measure the change in WERs when using different acoustic models trained with the additional data. We only trained the ML-SI and ML-HLDA-SAT models in this sequence. The additional data is added incrementally to the baseline 141-hour Hub4 acoustic training data set. The testing material used in the experiments is the EARS BN 2003 development test set [10]. The test set comprises the first 30 minutes of the six shows selected from the TDT4 subset.

5.1. Effects of Adding TDT Data

Table 3 exhibits four sets of WERs when using four different sets of acoustic models trained on increasing amounts of data. The first two columns list the data sets and their total amounts of training data in hour. The last three columns show the WERs at each decoding stage: speaker-independent decoding (SI), first adapted decoding (Adapt-1), and second adapted decoding (Adapt-2). The system results are the Adapt-2 WERs. In the first row, we established the baseline result using the 141 hours of Hub4 data *carefully* transcribed by human.

When doubling the data, from 141 hours to 297 hours, the WER went down from 12.7% to 12.0%, or 0.7% absolute reduction. Doubling the amount of data again, from 297 hours to 602 hours, we obtained another 0.6% absolute gain to bring the WER down to 11.4%. Adding another 240 hours produced another 0.5% reduction. Overall, by adding 702 hours to the baseline 141 hours, the WER went down from 12.7% to 10.9%, or 14.2% relative reduction.

It is encouraging that the WERs keeps decreasing with more train-

Data Set	Hrs	SI	Adapt-1	Adapt-2
h4 (baseline)	141	17.2	13.0	12.7
h4+tdt4	297	15.4	12.2	12.0
h4+tdt4+tdt2	602	14.7	11.6	11.4
h4+tdt4+tdt2+tdt3	843	14.5	11.2	10.9

Table 3: Comparison of WERs when adding more data

ing data. It is also interesting to see that the light supervision method seems effective. However, we might never be able to quantify its effectiveness unless we have all 1400 hours of the TDT data transcribed carefully by human. Another noteworthy point to make is that when having *matched* data in training and testing, the performance of the SI models improved significantly (from 17.2% to 14.5%, or 15.7% relative).

If we look at the per-show WERs in Table 4, we can see that the relative reduction seems consistent for all shows, ranging from 10.9% to 18.6%. However, there was little or no gains for the MSN and NBC shows when adding the TDT2 data (9.0 and 10.0 vs. 9.0 and 10.2). These two sources are not part of the TDT2 subset.

Hrs	ABC	CNN	MSN	NBC	PRI	VOA	All
141	11.4	18.6	9.8	11.5	9.7	15.6	12.7
297	10.8	18.0	9.0	10.2	9.2	15.0	12.0
602	10.3	16.7	9.0	10.0	8.8	14.0	11.4
843	9.6	16.1	8.4	9.7	7.9	13.9	10.9

Table 4: Comparison of per-show WERs when adding more data

5.2. Model Scalability

Since Byblos's system parameters are mostly data-driven, we intentionally kept the same thresholds as those used in the 141-hour baseline to let the systems decide the number of free parameters to use as the amount of data increased. As shown in Table 5, when the amount of data increased, the number of mixture densities (in Column 4), hence the number of Gaussians (in Column 5), increased linearly. There is another research effort at BBN to determine a more optimal *growth* function for this matter, since we are expecting to process 5000 or even 10000 hours of new data in a near future.

Data Set	Hrs	spkrs	cbks	gauss
h4	141h	7k	6k	164k
h4+tdt4	297h	12k	13k	354k
h4+tdt4+tdt2	602h	23k	26k	720k
h4+tdt4+tdt2+tdt3	843h	31k	34k	983k

Table 5: System parameters for different training sets

5.3. MMI-HLDA-SAT Results

When we used the MMI models in place of the HLDA-SAT models, we obtained another set of results shown in Table 6. The WER decreased from 12.1% to 10.5%, or 13.2% relative. This relative reduction is lower than the 14.2% when using HLDA-SAT models. This could be due to the fact that there are too many Gaussians in the 843-hour system, since our experience showed that MMI models are more effective when the system uses smaller number of parameters. Or this could be due to the transcription errors in the additional 702 hours of data derived automatically, since MMI training is known to be quite sensitive to the truth reference.

Data Set	Hrs	SI	Adapt-1	Adapt-2
h4 (baseline)	141	17.2	12.2	12.1
h4+tdt4+td2+tdt3	843	14.5	10.6	10.5

Table 6: Comparison of WERs when using MMI models

6. CONCLUSION

We have just presented a new *light supervision* method to automatically acquire additional acoustic training data from broadcast news audio having corresponding closed-caption transcripts. The method seems effective and able to select 72% of the captioned data of the TDT corpus. Biased language models play a key role in this method. When adding 702 hours of selected data to the baseline 141-hour training data set, we achieved a 13% relative error rate reduction.

References

1. L. Lamel, J.-L. Gauvain and G. Adda, "Investigating lightly supervised acoustic model training," *Proc. ICASSP*, Salt Lake City, May 2001.
2. P. Nguyen, "Large corpus experiments for broadcast news recognition," *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003.
3. L. Nguyen, et al., "The 1999 BBN Byblos 10xRT broadcast news transcription system," *Proc. NIST 2000 Speech Transcription Workshop*, Maryland, May 2000.
4. L. Nguyen and R. Schwartz, "Efficient 2-pass N-Best decoder," *Proc. EuroSpeech*, Rhodes, Greece, Sep. 1997, pp. 167-170.
5. L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, 38, pp. 213-230, 2002.
6. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, April 1990.
7. S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," to appear in *IEEE ASRU Workshop*, St. Thomas, Nov. 2003.
8. M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, 12, pp. 75-98, 1998.
9. P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition", *Computer Speech and Language*, 16, pp.25-47, 2002.
10. R. Schwartz, et al., "Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system," Submitted to *ICASSP*, Montreal, Canada, May 2004.