UNSUPERVISED TRAINING ON LARGE AMOUNTS OF BROADCAST NEWS DATA

Jeff Ma, Spyros Matsoukas, Owen Kimball, Richard Schwartz

BBN Technologies

ABSTRACT

This paper presents our recent effort that aims at improving our Arabic Broadcast News (BN) recognition system by using thousands of hours of un-transcribed Arabic audio in the way of unsupervised training. Unsupervised training is first carried out on the 1,900-hour English Topic Detection and Tracking (TDT) data and is compared with the lightly-supervised training method that we have used for the DARPA EARS evaluations. The comparison shows that unsupervised training produces a 21.7% relative reduction in word error rate (WER), which is comparable to the gain obtained with light supervision methods. The same unsupervised training strategy carried out on a similar amount of Arabic BN data produces an 11.6% relative gain. The gain, though considerable, is substantially smaller than what is observed on the English data. Our initial work towards understanding the reasons for this difference is also described.

1. INTRODUCTION

Unsupervised training has been shown [1, 2] to be an effective method to train acoustic models in the situation where un-transcribed audio data is available. In [2], the unsupervised training method is able to reduce the word error rate (WER) significantly even starting with a very limited amount of transcribed – or seed – data (10 minutes). Our interests here are to run unsupervised training on large amounts of data, starting from a sizable amount of transcribed data. For both English and Arabic BN, thousands of hours of audio data has been recorded and released by LDC, such as the 1900-hour TDT English BN data and a similar amount of newly released Arabic BN data – recorded from a variety of Arabic sources. The 1900-hour English TDT data is closed captioned, consisting of 4 parts, TDT2 (633 hours), TDT3 (475 hours), TDT4 (294 hours) and TDT4 extra (465 hours. The availability of closed captions on this data has led to the development of lightly supervised training methods to improve the performance of English BN recognition systems [3, 4, 5, 6]. The lightly supervised training has made a substantial contribution to the achievements that were made during the EARS RT04 evaluation by those participants [8].

It is certainly desirable to use the lightly supervised training method to take advantage of large amounts of training data in other languages, besides English. One such language of importance is Arabic. However, the newly released Arabic BN audio data has no closed captions. Thus, to utilize the large amount of audio data, the unsupervised training method needs to be adopted.

Instead of directly exploring unsupervised training on the Arabic BN data, we decided to develop the training procedure on the 1,900-hour TDT English BN data first, so that we can compare it with the lightly supervised training[1] approach.

This paper is organized as follows: Section 2 reports results with lightly-supervised and unsupervised training on the 1,900-hour English data; Section 3 addresses the unsupervised training experiments carried out on a similar amount of Arabic data, and reports on initial analysis of the results; Section 4 concludes this paper and discusses future research directions.

2. UNSUPERVISED TRAINING ON ENGLISH BN DATA

The carefully transcribed 140 hours of Hub4 data were chosen as the seed data in our experiments. Models were trained on the seed data and were used to decode the TDT data. Before the decoding, the data was first automatically segmented by using our segmentation tools, which remove long-silence and non-speech regions. 1,700 out of the 1,900 hours remained after the segmentation. All the experiments presented in this section started with this 1,700-hour data.

2.1. Revisiting light supervision

In the lightly supervised training we carried out for the RT04 evaluation [5], Maximum Likelihood (ML) models were used to decode the TDT data. We revisited the lightly-supervised training by switching to MMI models for decoding the 1,700-hour data. The MMI models trained on the 140-hour seed data achieve a WER 13.5% on the Hub4 Dev04 (h4d04) test set after unsupervised speaker adaptation. The speed of the baseline decoding experiment was approximately 10xRT, too slow to run on the 1,900-hour TDT data. We therefore took an effort to speed up the decoding. Specifically, we removed the I/O intensive 4-gram LM rescoring and increased various pruning beam thresholds. The removal of the 4-gram LM rescoring degrades the performance by 0.5% absolute, and the increase of the pruning causes another 0.3% absolute degradation. Thus, the total loss is 0.8% absolute – or 6.0% relative. But, after these changes, the decoding ran at 3.8xRT, which was acceptable to us. We don't believe that the 0.8% degradation makes a big difference in the data selection, and hence we used this faster configuration to decode the 1,700-hour data.

Following what was done in our RT04 evaluation, in this set of experiments, biased language models (LMs) were used in the decoding of the TDT data. The biased LMs were trained by giving the closed captions a high weight (set to 4.0, a weight of 1.0 was used for other data). **Table 1** lists the experiments we ran in this

[1] In [6], a similar comparison was made, but only on the TDT2 subset.

scenario, where models trained under the ML criterion (ML models) and models trained by using the speaker adaptive training method under the maximum mutual information criterion (MMI-SAT models) were used in the un-adapted and the adapted decoding passes, respectively.

Expt	Train data (hours)	Model size	Un-adapted WER	Adapted WER
B	140	230K	17.5	13.5
L1	140+980	603K	13.2	9.8
L2	140+1700	872K	13.4	-

Table 1. Lightly supervised training on the English BN 1,700-hour TDT data with biased LMs (WERs on h4d04; model sizes given in number of Gaussians).

Experiment "B" is the baseline model trained on the seed data. Experiment "L1" selected from the 1700-hour hypotheses all the word phrases that matched the corresponding closed captions (as in [5]), adding a total of 980 hours of data to the training. It produced a 4.3% absolute gain – a significant gain – in the un-adapted decoding pass, compared to the baseline. In Experiment "L2", no hypothesis selection was performed; all the 1,700-hour data was added to the acoustic training. We can see that "L2" degrades the recognition accuracy by only 0.2% absolute compared to "L1". This implies that the gain from filtering hypotheses using the closed captions is not significant. Compared to the baseline, "L1" yields a 24.6% relative gain. **Table 1** also lists the adapted decoding results for Experiment "B" and "L1". The relative gain (27.4%) still remains after speaker adaptation.

2.2. Unsupervised training

In this unsupervised training scenario, we assume that all the TDT closed captions are not available, and, consequently, neither can we decode the 1,700-hour data with the biased LMs, nor can we use closed-caption matches in the data selection. We, therefore, decoded the 1,700-hour data using an unbiased LM, trained by excluding all the TDT closed captions. This LM change causes 4~5% absolute degradation in WER (20% relative), measured against the closed captions. Without guidance from the closed captions, we have compared three approaches to select data. One is to select all the data blindly, and the other two are to select data based on word confidence scores, computed from N-best hypothesis lists.

The first confidence-based method is to select utterances if their confidence scores are above a threshold (referred to as confidence-based sentence selection in what follows). In this method, the utterance confidence is a weighted sum of word confidences, and it is computed according to:

$$conf\ \{S\} = \sum_{i=1}^{N} conf\ \{W_i\} * len\{W_i\} / \sum_{i=1}^{N} len\{W_i\} \qquad (1)$$

where N is the number of words in the utterance, and $conf\{W_i\}$ and $len\{W_i\}$ are the confidence and duration of word W_i, respectively.

The second confidence-based method is to select word phrases if all the word confidence scores in the phrases are above a threshold (referred to as confidence-based word-phrase selection method).

Expt	Data selection (method, thresh)	Train data (hours)	Un-adapted WER
U1	Select all, 0.0	140+1700	13.8
U2	select utts, 0.8	140+1060	14.1
U3	Select word-phrases, 0.8	140+1036	14.2
U4	Select utts, 0.5	140+1479	13.7

Table 2. Unsupervised training on the English BN 1,700-hour TDT data with unbiased LMs (WER measured on h4d04 test set).

A comparison between these three methods is shown in Table 2, where all experiments used the ML models to get the un-adapted WER. Experiment "U1" selected all the data blindly. Recall that Experiment "L2" in **Table 1** selected all the data as well, but used the biased LMs in the data selection. Comparing these two experiments, we can see that the switch from the biased LM to the un-biased LM in the data selection causes 0.4% absolute degradation in the un-adapted decoding pass. Experiment "U2" tried the confidence-based utterance selection and "U3" the confidence-based word-phrase selection. They both set the threshold to 0.8 and selected similar amounts of data (rejected more than 30% of the data). We see that the utterance selection outperforms the word-phrase selection slightly (0.1% absolute). But, they both degrade, compared to "U1", which selected all data. Experiment "U4" repeated "U2" but with a lower threshold, 0.5, thereby rejecting less data (14.3% is rejected). This experiment produced a minor WER reduction (0.1% absolute) compared to "U1", recovering the loss due to the high-percentage rejection, observed in "U2" and "U3".

In all the above unsupervised experiments, the selected data and the seed data were weighted equally in the acoustic training. In general, giving the same confidence to the selected data as to the seed data may not be the best choice, because the seed data is well transcribed. Therefore, we repeated Experiment "U4" listed in Table 2 but gave the utterances in the seed data weights that are equal to 1 and the utterances in the selected data weights that are equal to their confidences computed according to Eqn (1). The utterance scores are between 0 and 1, so the seed data is weighted more heavily. We found that such a weighting yields a minor gain (0.1% absolute).

We also found that the confidence-based data selection is able to exclude the majority of the commercials present in the BN audio. Since some commercials still remained, we tried the commercial-removal algorithm used in [7]. The algorithm assumes that the commercials repeat but the news or stories don't, and detects repeats based on the Arithmetic Harmonic Sphericity (AHS) distance. Different from [7], we searched and removed repeats after the data was selected. It removed 2% of the data, with no impacts on the WER.

The best un-adapted WER in Table 1 is 13.2% ("L2"), and the best one in Table 2 is 13.7% ("U4"). So the switch from lightly supervised training with biased LMs to unsupervised training with unbiased LMs causes only 0.5% absolute loss. Compared with the baseline ("B" in Table 1, WER 17.5%), the WER 13.7% is 3.8% absolute lower, and the relative improvement is 21.7%. This still

is a significant gain. So, it assures us to use the unsupervised training in situations where no closed captions are available, such as Arabic BN.

3. UNSUPERVISED TRAINING ON ARABIC BN DATA

Unsupervised training experiments we carried out on Arabic BN made use of 1,858 hours of un-transcribed audio data, comprised of 6 different sources: Al Jazeera, LBC, Dubai, SSTV, EDTV, and Al Alam. As in the case of English BN, the audio was first automatically segmented, retaining 1,570 hours of data. All unsupervised training experiments reported in this section were carried out on this 1,570-hour data set.

We used a total of 150 hours of Arabic BN speech as the seed data, consisting of 28 hours of carefully transcribed data (from FBIS), 67 hours of quickly transcribed TDT4 data (selected automatically via light supervision), and 55 hours of quickly transcribed in-house data (also selected automatically via light supervision). The unbiased LMs were estimated from a variety of text sources (Arabic Gigaword corpus, Al Jezeera web data, etc.), totaling 261M words. On the Arabic BN dev04 test set (h4ad04), the MMI models trained on the 150-hour seed data achieve WERs of 17.4% and 14.8% in the un-adapted and the adapted decoding passes, respectively. This set of MMI models along with the unbiased LMs were then used to decode the 1,570-hour data.

3.1. Training procedure and initial results

The unsupervised training method that performed the best on the 1,700-hour English data (Experiment "U4" listed in Table 2) was first repeated on the Arabic data. It is denoted as Experiment "AU2" in Table 3, where again ML models were used in the un-adapted decoding pass and MMI models in the adapted decoding pass. With the threshold 0.5 for the confidence-based utterance selection, 1,365 hours out of the 1,570-hour data were selected. This experiment produces 1.6% absolute (or 8.8% relative) gain compared to the baseline model trained on the 150-hour data (Experiment "AU1" listed in the same table). This gain is much smaller than the 3.8% absolute (or 21.7% relative) gain we got on the English data. To investigate the reasons, we repeated "AU2" but with a higher threshold, 0.85 (Experiment "AU3" in Table 3). Recall that on English BN, increasing the threshold in the confidence-based utterance selection hurts recognition accuracy. On Arabic BN, however, the results are reversed. Using a higher threshold improves performance by 0.2% absolute, even though much less data, 488 hours, is selected. This result implies that, unlike English, the unsupervised Arabic BN data contains hypotheses with decent confidence scores, but they either are mismatched to the testing data or have high WERs and hurt the performance of the retrained acoustic models.

More efforts were then taken to enlarge the gain. In general, a better acoustic model is able to exclude more garbage data when used in the data selection. Hence, we have tried adding data incrementally, a practice that is commonly adopted in the unsupervised training. In this scenario, data is typically divided into subsets, and the subsets are added into the acoustic training one by one. In this way, the model gets better after each loop, resulting in improved decoding quality on the remaining data[2]. We tried this method by splitting the Arabic data into halves. "AU4" in Table 3 represents the experiment that added the first half (755 hours were selected from the first half of the data, using a confidence threshold of 0.5).

Expt	Thres-hold	Train data (hours)	Model size	Un-adaptd WER	adapted WER
AU1	-	150	222K	18.1	14.8
AU2	0.5	150+1365	748K	16.5	-
AU3	0.85	150+488	388K	16.3	13.7
AU4	0.5	150+755	451K	16.8	14.1

Table 3. Unsupervised training results on the 1,570-hour Arabic BN data (WERs measured on h4ad04 test set).

The performance of "AU4" is inferior (0.5% absolute worse in the un-adapted decoding and 0.4% absolute worse in adapted decoding) to that of "AU3". It has inspired us of a new procedure, where the incremental training is carried out based on decreasing confidence thresholds in the data selection step. The new procedure is as follows:

1. Use a model, M_i, to decode all the data.
2. Select data with a threshold, T_i.
3. Train a new acoustic model, M_{i+1}, with the data selected in step 2 added to the seed data.
4. If no further improvement is obtained, stop; otherwise, choose T_{i+1} to be a value less than T_i.
5. Set $i = i + 1$, and go to step 1.

$i = 1,2,3,...$, and M_1 is the seed model. It is reasonable to choose the threshold such that $T_{i+1} < T_i$ because the model becomes more trustable after each loop.

In our experiments, T_1 was chosen to be 0.85, which is experiment "AU3" in Table **3**. A new MMI model, M_2, was trained by adding the 488-hour data, and it achieved a WER of 13.7% after the speaker adaptation. The WER achieved by the seed MMI model, M_1, is 14.8% after the speaker adaptation. So the new MMI model outperforms the old one by 1.1% absolute. Next, the new MMI model was used to decode the data again, and a lower threshold, 0.75, was used to re-select data. The re-selection picked 922 hours of data. The new ML model – a prerequisite for the new MMI training – trained by adding this 922-hour data to the seed data achieves a WER 16.0% in the un-adapted decoding pass, which is 0.3% absolute lower compared to "AU3" in Table 3. So, this form of incremental training yields 0.3% absolute gain over the non-incremental case. Since this 0.3% gain is insignificant, we have stopped running more iterations. Overall, this new incremental procedure increases the relative improvement against the baseline to 11.6%. This improvement, however, is still much smaller than the 21.7% we obtained on the English data.

[2] Because the unsupervised training on the English data performed well, we did not carry out such experiments there.

3.2. Improving the data selection

The initial results of unsupervised training on Arabic BN suggest that data selection is an important step in the unsupervised training procedure, contrary to previous results we got on English BN. It's possible that there is a data quality issue in the Arabic corpus. To see if this is true, we devoted some time to manually inspect a small subset of the data. First, we checked some of the automatic segmentation results. We found that 11 out of 485 randomly selected segments were either pure music or speech with music background and the remaining segments were pure speech. So, only 2.3% of the automatic segments were problematic. It means that our automatic segmentation has worked well to remove non-speech regions. Second, we picked 6 episodes randomly – about one from each different source – and listened to them. The result is quite surprising. Only about 48% (less than half) of the data in the 6 episodes is news speech, and the remaining 52% is non-news speech, which is either music or speech with background music or drama dialogues. One episode has no news speech at all. Apparently, better algorithms are needed to distinguish news speeches from non-news speeches.

As our first effort in this direction, we have computed the LM perplexity for each of the 6 episodes (on their decoding hypotheses), and found that episodes having less percentage of news data have higher perplexity values and the episode without any news speech has an extremely high perplexity. Based on this observation, we have run experiments to simply remove episodes that have LM perplexities higher than a threshold. The threshold is source-dependent. For each source, the average perplexity is computed over all episodes from that source, and the source-dependent threshold is set to the source-dependent average perplexity multiplied by a scalar. Two such experiments are listed in Table 4. The experiments were run on top of experiment "AU3" listed in Table 3, i.e. they removed the episodes from the data selected in "AU3". When the removal thresholds are set to the average perplexities multiplied by 2.0 (experiment "AU31" in Table 4), about 5% of the data is removed (from 1365 hours down to 1305 hours), resulting in 0.3% absolute gain (compared to "AU3"). More aggressively, when we set the removal thresholds to the average perplexities multiplied by 1.5 (experiment "AU32"), about 10% data is removed, and it starts to hurt the performance.

Expt	Epsd-rmvl threshold	Train data (hours)	Model size	Un-adaptd WER
AU3	No	150+1365	748K	16.5
AU31	2.0*avg_ppl	150+1305	721K	16.2
AU32	1.5*avg_ppl	150+1227	685K	16.6

Table 4. Experiments with episode removals based on LM perplexities (WER on h4ad04).

The 0.3% gain from the use of this perplexity-based episode removing method is not significant. We are currently looking for other ways to improve the data selection.

4. CONCLUSIONS AND FUTURE WORK

This paper was described the work that we have done so far to improve our Arabic BN recognition system by using unsupervised training. On 1,900 hours of English TDT data, a comparison between unsupervised training and the lightly supervised training method we used for the EARS evaluations shows that the switch from the lightly supervised training with biased LMs to the unsupervised training with unbiased LMs only causes a minor loss (0.5% absolute). Unsupervised training on the TDT data yields a 21.7% relative WER reduction. The same unsupervised training strategy, carried out on the 1,858-hour Arabic BN data, produced a 8.8% relative gain. This gain is substantially smaller than what is obtained on the similar amount of English data. A new incremental training procedure was developed, based on decreasing confidence thresholds in the data selection step. Although this new method was able to increase the gain to 11.6% relative, the performance of unsupervised training on Arabic BN is still worse than what was observed on the English data. Preliminary inspection of parts of the Arabic data has revealed that less than half of the data is pure news speech. An episode-removing method based on LM perplexities, as our initial effort in this direction, turns out a limited improvement So, one of our future efforts will be to look for better algorithms to exclude non-news speech regions. Other efforts will include the use of cross-system or system-combination results in the unsupervised training.

5. ACKNOWLEDGMENTS

We would like to express our thanks to B. Xiang for training some of the LMs used in our experiments and to S. Wang for manually checking the automatic segments and marking the non-news speech regions.

6. REFERENCES

[1] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments", Proc. Eurospeech 99, pp 2725-2728, September 1999.
[2] L. Lamel, J. L. Gauvain, G. Adda, "Unsupervised acoustic model training", Proc. ICASSP 2002, pp 877-880, 2002.
[3] L. Lamel, J.L. Gauvain, G. Adda, "Investigating lightly supervised acoustic model training", Proc. ICASSP 2001, pp 477-480, May 2001.
[4] L. Nugyen, B. Xiang, "Light supervision in acoustic training", Proc. ICASSP 2004, pp 185-188, 2004.
[5] L. Nguyen, B. Xiang, M. Afify, etc., "The BBN RT04 English broadcast news transcription system", Eurospeech05, Lisbon, Sept. 2005.
[6] H.Y. Chan, P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", ICASSP2004, pp 737-740, 2004.
[7] S. E. Johnson, P. C. Woodland, "A method for direct audio search with applications to indexing and retrieval", ICASSP 2000, pp 1427-1430, 2000.
[8] The proceeding of "The EARS RT04 Evaluation Workshop", www.sainc.com/richtrans2004, Palisades, NY, Nov. 2004.