Full Utilization of Closed-captions in Broadcast News Recognition

Meng Meng, Shijin Wang, Jiaen Liang, Peng Ding, Bo Xu.

Hi-Tech Innovation Center, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080
{mmeng, sjwang, jeliang, pding, xubo}@hitic.ia.ac.cn

Abstract. Lightly supervised acoustic model training has been recognized as an effective way to improve acoustic model training for broadcast news recognition. In this paper, a new approach is introduced to both fully utilize the un-transcribed data by using closed captions as transcripts and to select more informative data for acoustic model training. We will show that this approach is superior to regular method, which filters data only based on matching degree of closed-captions and ASR results without considering the effectiveness of data. By the way, an approximately correct transcription for manual amendment is obtained by this approach, which can reduce manual effort enormously for detailed annotation.

Keywords: Lightly supervised acoustic model training, closed-caption, ASR.

1 Introduction

For acoustic model training, collecting a large number of transcribed acoustic training data is an expensive process in terms of both manpower and time. One the other hand, massive amount of unlabeled raw data is relatively easy to obtain from radio and television broadcasts, etc. Thus there exists the need for an automatic learning procedure that can use both labeled and unlabeled data for model training. This kind of training approach is usually referred as lightly supervised or unsupervised acoustic model training. The main difference between them is that whether closed captions are afforded.

The closed-captions are a close, but not exact transcription of what is being spoken. Usually they are only coarsely time-aligned with the audio signal and may contain many transcription errors. Though much less precise than detailed manual transcriptions, they are always available in sufficient amounts. Compared to actual speech transcriptions, closed captions lack a lot of features which make transcriptions valuable:

- The captions contain little or no time information, which makes them hard to convenient incorporated into conventional acoustic model training process.
- Insertions, deletions and changes in the word order are quite common. NIST reported the disagreement between the closed-captions and the manual transcripts to be on the order of 12% on 10 hours of TDT2 data.
The closed-captions do not have any indicators of non-speech events or disfluencies, such as breath noise, filler words, or fragments.

Background music and transmission noise are all marked in transcriptions but not in closed captions.

The identity of the speaker is better preserved in transcriptions where notes about speaker turns and gender, age, etc.

Therefore the closed-captions have to be refined with respect to all the above mentioned problems. In order to train acoustic models on audio data that only have closed-captions, some preparation procedure should be taken in general:

- Firstly, the audio data needs to be partitioned into homogeneous segments, automatically producing some of the missing information that is missing in the closed-captions, such as speaker, gender and bandwidth.
- Secondly, for the reasons discussed in the previous paragraph, the closed captions have to be aligned to the actual audio signal, which is typically done by a segment tool implemented by some dynamic programming approach.
- Afterwards the closed-captions are used to filter the hypothesized transcriptions. Transcriptions that do not match the closed captions are considered incorrect and therefore discarded.
- The resulting training corpus is often a quarter to a third smaller than the raw data. They are used to training acoustic model finally.

Closed captions, when available, are usually used as a filter to separate the usable data from the left [2], because they are reasonable approximation of transcriptions. Lamel et al. [4][5] used closed captions instead of confidence measure to filter the hypothesized transcriptions. They showed that the use of filtering with the closed-captions, which is essentially a perfect confidence measure, was found to slightly reduce the word error rate. An implicit assumption is that it is extremely unlikely that the recognizer and the closed-caption both have the same error rate.

Confidence measures and closed-captions are commonly used to decide what kind of data is more suitable for lightly supervised training. However, researchers have diverse opinions upon the effectiveness of using data with high confidence score or entire matching. For example, [7][6] suggests that this kind of data can’t add substantial new information to the existing recognizer, while [3] shows that the good performance is mainly due to the contribution of the large size of data without regarding to which type it is. [1] combine active and semi-supervised learning using selectively sampled and automatically labeled data for spoken language understanding. This method enables them to exploit all collected data and alleviates the data imbalance problem caused by employing only active or semi-supervised learning.

The closed-captions are inclined to be used as filter to select the scripts that appear accurate in the papers mentioned above. In this paper, closed-captions act not just as filter, but substitute as the transcriptions directly. The matching degree of closed-captions with hypothesized transcriptions is used as the rule to determine which segments are reserved as transcriptions and which are not.

Especially for broadcasting news data, it is easy to obtain corresponding closed-captions from the news website. Generally, these closed-captions are exact enough as what is being spoken except for some segments where the closed-captions are missing. The mismatching is usually located at the places where the acoustic conditions are bad such as music background, nonnative speaker, noisy environment, etc. Selecting
data, whose closed-captions are matched with hypothesized transcriptions not bad to training AM, is a rather suitable method to utilize closed-captions in lightly supervised training.

The general procedure of lightly supervised acoustic model training by the use of closed-caption is as follows.
1) Construct a language model (LM) using only the closed-caption data. Then this closed-caption LM was interpolated with a general BN LM using interpolation weights heavily weighted towards the closed-caption LM. This yielded a “biased” language model.
2) Recognize the audio data using an existing acoustic model and the biased LM trained in (1).
3) Optionally post-process the data. Replace hypothesized transcriptions by closed-captions with time alignment information to the signal.
4) Use the selected segments where the recognition output from (2) is consistent, to some level, with the closed-captions for acoustic model training with the closed-captions as transcriptions.

In this paper, the (1), (2) steps are skipped over: the closed-captions and hypothesized transcriptions which released in LDC TDT4 corpora are all ready-made in the following experiments. The next section presents the methods of utilize closed-captions novelly for lightly supervised training, followed by a description of the data corpora and the configuration of our transcription system used in this paper. The experimental results and discussion are given in Section 4.

2 Closed-captions as Transcript

The algorithm which is based on dynamic programming approach tries to match closed-captions with the result of ASR, and imparts the time mark information to the sentences in the closed-captions. Then these sentences with time marks are used to training AM. The basic assumption of the algorithm is that the closed-captions are correct in the rough. Further more, the Character Error Rate (CER) of each sentence which is calculated during DP procedure can be used as judgment for data selection. Other than only be used as filter to select more accurate scripts, closed-captions are utilized farther.

2.1 Revised DP algorithm

Firstly, closed-captions are marked with time stamp based on paragraph. There are usually many Arabic numerals in closed-captions which must be turned into characters in accordance with their pronunciations as far as possible. Then the closed-captions are partitioned into sentences with integrated meaning. On the other hand, the ASR result contains multi-character words together with their start times, durations and speaker clusters is obtained.

In order to improve the matching degree between closed-captions and ASR results, the basic DP algorithm is updated. Firstly, the matching units are characters instead of words, and their pronunciations are also taken into account. In the algorithm, Sim-
substitution is introduced besides Match, Insertion, Deletion and Substitution accustomed. It is used when the pronunciations of two characters are similar, and its cost which depends on this similarity has the value between costs of Match and Substitution. This skill enhanced the compactness and veracity of matching performance. When DP matching uses words as comparison unit, the difference of the number of characters they include will affect the matching just followed if mismatch occurred, even though the mismatch diversity on characters is tiny. It is proved that matching based on characters and the introduction of Sim-substitution improved algorithm’s robustness.

Compared with the ASR result, closed-captions, as approximation of actual transcriptions, often lose some sentences in general. In these places, the performance of DP procedure will extremely impaired if the recognition is not good enough. To resolve this problem, the cost of Insertion at the boundary of two sentences should be tuned in order to increase the probability of inserting mismatch segments rightly in between two whole sentences. It is useful to cope with the case that many sentences are absent in closed-captions.

In fact, where the matching of closed-captions and ASR result is farfetched, usually the ASR’s performance is unsatisfactory. Even then, these segments can be secluded from their surrounding effectually because their durations are not too long and it is usually very suitable just foreside and offside. For this reason, the data of these segments can be extracted fitly to be used after some other disposal.

Finally, in connection with the foregoing efforts, the time stamps in the ASR result are contacted to the whole sentences of closed-captions. The modus operandi of other institution is to select only parts of data that is credible enough on the result of the matching produce. The rule of selective is various, but we think that the sentences in closed caption are just the correct transcriptions on the whole. Especially for most situations, the closed caption is the same as the news release for the broadcasting news speech.

At the same time, the CER of each sentence is calculated regarding closed-captions as reference. By the way, the Sim-substitution is taken into account during the computation. We use these CERs for the following data filtering.

### 2.2 Farther utilization

Though we think that the sentences of closed-captions is better than the ASR result as transcription for acoustic model training in BN task, it is unavoidable that there are many errors between closed-captions and what is being spoken actually. Also there are time marking errors in the matching result which is built above. So we need to move a step further.

The CER of each sentence gives expression to performance of automatic recognizing system on it. It also can make out if it’s corresponding data are effective to enhance the performance of this system. Base on this opinion, we can use it as reference to select the data contains more information for training the AM.
3 Experiment Settings

3.1 Dataset

The acoustic training data consist of about 24 hours broadcast news data in Mandarin with accurate time-aligned transcriptions (1997 Hub4-Mandarin) and about 166 hours of data from the TDT4 corpus distributed by LDC with closed captions. The Hub4 data come from 3 sources: VOA, CCTV and KAZN-AM. The TDT4 data come from 5 sources: China Central Television (CTV), China National Radio (CNR), Voice of America - Mandarin Chinese news programs (VOA); China Broadcasting System (CBS) and China Television System (CTS), which were collected from October 2000 to January 2001.

The actual sampling rate of CTS and CBS source is 8k Hz and the character error rate (CER) is particularly high. So they are moved away from the experiment data and the reserved training data corpus (3 sources of mainland style) is named as TDT4slt3.

The closed-captions and the result of ASR used in this article is just what LDC released, src_sgm and asr. src_sgm contains original source text data from which reference texts are derived, in an SGML markup format. asr is output of the BBN automatic transcription engine in multi-character words.

After match of ASR and SRC use the algorithm which introduced in section 2, the average CER of each data source of the TDT4 are list as follows: 47.3% (CBS), 70.4% (CTS), 16.1% (VOA), 15.3% (CNR) and 17.0% (CTV). All data is partitioned into 7 bins base on the sentences’ CER. The size of each block in separate data source is list in Table 1.

<table>
<thead>
<tr>
<th>Bin</th>
<th>%CER</th>
<th>CNR</th>
<th>CTV</th>
<th>VOA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin0</td>
<td>0−10</td>
<td>19.11</td>
<td>13.67</td>
<td>29.44</td>
<td>62.22</td>
</tr>
<tr>
<td>Bin1</td>
<td>10−20</td>
<td>3.48</td>
<td>4.19</td>
<td>8.09</td>
<td>15.76</td>
</tr>
<tr>
<td>Bin2</td>
<td>20−30</td>
<td>1.65</td>
<td>2.06</td>
<td>3.22</td>
<td>6.93</td>
</tr>
<tr>
<td>Bin3</td>
<td>30−45</td>
<td>1.26</td>
<td>1.55</td>
<td>2.00</td>
<td>4.81</td>
</tr>
<tr>
<td>Bin4</td>
<td>45−60</td>
<td>0.82</td>
<td>1.00</td>
<td>1.06</td>
<td>2.88</td>
</tr>
<tr>
<td>Bin5</td>
<td>60−80</td>
<td>0.89</td>
<td>0.83</td>
<td>0.92</td>
<td>2.64</td>
</tr>
<tr>
<td>Bin6</td>
<td>80−100</td>
<td>1.22</td>
<td>0.72</td>
<td>2.09</td>
<td>4.03</td>
</tr>
<tr>
<td>Total</td>
<td>0−100</td>
<td>28.43</td>
<td>24.02</td>
<td>46.82</td>
<td>99.27</td>
</tr>
</tbody>
</table>

3.2 System configuration

In this article there are two test set, TSET1 and TSET2, which separately come from two mandarin TV stations. Each test set is recorded recently and comprises with both the broadcast news and the talk shows on the ratio of 3:2 to present comprehensively the BN program types. TSET1 is more difficult than TSET2 because of more
interview outdoor and nonnative speakers’ speech existing in it. Each set is about 2 hours.

The acoustic models used in following experiments are all ML Trained without any optimization, in order to compare the performance of different means of lightly supervised training and speeds the training procedure as well. A 13-dimension PLP feature vector is computed for each frame, 1 pitch dimension is added and then converted to a 42-dimension acoustic feature vector by adding delta and delta-delta coefficients. The features are normalized by per utterance cepstral mean subtraction. In all AMs discussed in this section, 32 gaussians are estimated per feature per state.

The language models for this two test sets is built on 2 sources of texts: large amounts of newspaper and newswire texts, much smaller amounts of detailed BN transcriptions corresponding to the program which the test set comes from.

4 Experiments and Results

In the experiments followed, we compare the method above mentioned with the usually means. The effective of different type of data is examined.

4.1 Method comparison

In order to see the actual improvement obtained with active learning, we performed controlled experiments comparing our methods with ordinary means.

Table 2 gives the CER on two test set, using three different methods for lightly supervised AM training, with TDT4 closed-captions. Content in parenthesis is the size of data which took the hour as the unit. In fact, data used for training acoustic models is appreciably less than that, because there are some data is abandoned for failure of Viterbi forced alignment of the transcript and the signal. The first method uses the data whose recognition result is equal to the closed-captions. About 20% data is abandoned by reason of mismatching. The second method uses all data to training with the result of recognition as script. The third method, as mentioned above, uses about all data with the sentences in closed-captions instead of the ASR result as script for AM training. The result shows that our method performs best within three methods and using all data without regard to closed-captions is comparable to the ordinary means which only select data “correct”.

Table 2. Comparison of 3 methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>%CER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TSET1</td>
</tr>
<tr>
<td>Correctly recognized (80h)</td>
<td>35.28</td>
</tr>
<tr>
<td>All recognized (99h)</td>
<td>34.95</td>
</tr>
<tr>
<td>Script replaced by closed-captions (99h)</td>
<td>34.40</td>
</tr>
</tbody>
</table>
4.2 Data selection

The results in Table 2 indicate that only “correct” data may not be the best choice for lightly supervised training even though compared with use all data. We did next experiment to further study this question. In our method, if we select data considering sentences’ CER, can we arrive some better performance?

As shown in Table 1, data whose CER of sentences below 20%, is added up to about 80% of all data in 3 mainland style sources. These data is comparable with the data that correctly recognized, considering accuracy of script and the size of data. So we treat these data as base set. Data in higher CER bin in succession is added step by step. The performance of this group of data set is shown in Table 3.

When only using data whose CER below 20%, there are only tiny differentials at CER on test set, compared with using data which correctly recognized. With data in neighboring bin added, the CER is declined even though little size is increased relatively (Data whose sentences’ CER is between 20% and 30% is only about 8 hours). The trend last out until all data CER<60% is added. The performance is drop when any more data is superadded. A 3% relative reduction is achieved.

Table 3. Comparison of data selection rules.

<table>
<thead>
<tr>
<th>Data selection rule</th>
<th>%CER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TSET1</td>
</tr>
<tr>
<td>CER&lt;20% (78h)</td>
<td>34.90</td>
</tr>
<tr>
<td>CER&lt;30% (85h)</td>
<td>34.33</td>
</tr>
<tr>
<td>CER&lt;45% (90h)</td>
<td>34.16</td>
</tr>
<tr>
<td>CER&lt;60% (93h)</td>
<td>34.26</td>
</tr>
<tr>
<td>All sentences (99h)</td>
<td>34.40</td>
</tr>
<tr>
<td>CER&gt;60% amended (99h)</td>
<td>34.12</td>
</tr>
</tbody>
</table>

4.3 Training with Hub4 set

Table 4. Training together with Hub4 set.

<table>
<thead>
<tr>
<th>Base Data</th>
<th>Add data</th>
<th>%CER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TSET1</td>
</tr>
<tr>
<td>Hub4</td>
<td>none</td>
<td>36.42</td>
</tr>
<tr>
<td>Correctly recognized (80h)</td>
<td>34.88</td>
<td>26.05</td>
</tr>
<tr>
<td>All recognized (99h)</td>
<td>34.37</td>
<td>26.06</td>
</tr>
<tr>
<td>Script replaced by closed-caption (99h)</td>
<td>34.12</td>
<td>25.46</td>
</tr>
<tr>
<td>CER&lt;60% (93h)</td>
<td>33.95</td>
<td>25.33</td>
</tr>
<tr>
<td>CER&gt;60% amended (99h)</td>
<td>33.86</td>
<td>25.30</td>
</tr>
</tbody>
</table>
When training together TDT4slt3 with Hub4 set, the results under different conditions are shown in Table 4. The performance of these approaches is parallel with conclusions in above experiments. The best result comes from the data selected based on CER level, only abandon what matched worst. There is about 2% relative decrease compared to only use correctly recognized data for training. Unfortunately, the result on TSET2 is worse than only use TDT4slt3 for training. We think it is because the diversity of HUB4 and TSET2 is huge.

4.4 Closed-captions modified manually

Experiments above show that the data whose sentences’ CER>60% is of no help to improve the effectiveness of AM. So we next put these data to be modified manually. Fortunately, the size of this part of data is relatively small.

In the case of amendment, most of these sentences are found to be marked with error time stamp or mismatched with what being spoken severely. The following situations appear frequently:

1. There is music or other background noise which is distinct along with speech; 2. Overlapping with other speaker’s voice also occurred; 3. Within interview outdoor or through telephone, nonnative speakers or diverse sampling rate is ineluctable; 4. There are slip and amendment in speech; 5. The audio signal is distorted; 6. There are distinct errors in closed-captions such as repetition or upside down; 7. Foreign words are intermingled with Chinese; etc.

It is not time-consumeing to modify these sentences because they are only a few and have been changed into the format compatible with transcriber tools. Closed-captions modified manually were found to slightly reduce the character error rate which is shown in Table 3 and Table 4. But this approach should save the cost of time and manpower greatly for detailed manual annotation.

4.5 Discussion

When using only the data whose sentences’ CER<20% to train AM, many complex phenomena of acoustics are thrown out together with the data whose sentences’ CER are high. Even though the size of data in the bin 2 to bin 4 (CER between 20% and 60%) accounts for a smaller proportion of all (15%), they make an obvious contribution to the performance. It is because that, although the CER of this part of data is high, quiet a part of sentences are just the actual scripts and the relevant data contains many new and complex realistic environment information. When more data (CER above 60%) is added, the performance is worse instead. The reason, we think, is that there are many writing mistake in closed-captions and many time errors when align transcript to the signal within DP procedure. This part of data with scripts should be regarded as invalid data if no further processing be made.
5 Conclusions and Future Work

We have investigated the use of closed captions roundly instead of in some certain extent for train acoustic models in broadcast news recognition. It is shown that recognition results obtained with acoustic models trained on data whose transcripts are just corresponding closed-captions are overmatch to results with acoustic models trained on automatically annotated data filtered by closed-captions. Instead of selecting part of data which match closed-captions, our approach uses closed-captions as actual transcripts to the majority of data which have better automatically annotate result. The main advantage of this approach is that we can add some data which is not easy to be recognized correctly but contains more new information to train acoustic model. More useful data and more precise context can be used conduces to improve the performance of lightly supervised training.

The limit of this method is that some tiny errors in closed-captions can not be avoided; therefore the performance of this approach under the influence of precision of closed-captions. One possible way is combining confidence measure with closed-captions to revise these tiny errors. In addition, enhancing the flexibleness of Viterbi forced alignment administers to further amelioration. We hope to study these concerns in the next few months of experimentation.

References