AUTOMATIC TRANSCRIPTION VERIFICATION OF BROADCAST NEWS AND SIMILAR SPEECH CORPORA

Michael Pitz, Sirko Molau, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik VI
RWTH Aachen
Germany

ABSTRACT

In the last few years, the focus in ASR research has shifted from the recognition of clean read speech (i.e. WSJ) to the more challenging task of transcribing found speech like broadcast news (Hub-4 task) and telephone conversations (Switchboard). Available training corpora tend to become larger and more erroneous than before, as transcribing found speech is more difficult. In this paper we present a method to automatically detect faulty training scripts. Based on the Hub-4 task we will report on the efficiency of error detection with the proposed method and investigate the effect of both manually and automatically cleaned training corpora on the word error rate (WER) of the RWTH large vocabulary continuous speech recognition (LVCSR) system.
This work is a joint effort of the University of Technology (RWTH) and Philips Research Laboratories Aachen, Germany.

1. INTRODUCTION

The importance of automatic transcription verification was highlighted by the 1997 Hub-4 Broadcast News evaluation. A number of participating sites reported efforts to clean transcriptions of the training material hereby improving the quality of their acoustic models [1, 2]. Either the whole corpus was manually checked and corrected, or suspicious speech segments with bad scores were rejected during training. Our first tests with the RWTH LVCSR on the 1996 Hub-4 evaluation task supported this procedure. We obtained a reduction in WER (table 1) when training on a subset of manually corrected 46 hours compared to 76 hours of uncorrected data, released 1996 and 1997 from LDC. Even though the small subset contained only 60% of the available data, the WER decreased. This indicates that our preliminary system was rather sensitive to incorrect transcriptions.

1996/97 Hub-4 training corpus	size	WER
manually corrected	46h	36.7%
complete	76h	37.1%

Table 1: Recognition results on Hub-4 '96 eval. set, obtained with the *preliminary* RWTH LVCSR system trained on different training corpora. All WER reported in this paper are obtained by NIST scoring.

2. SELECTION CRITERIA

2.1. System Description

The error detection was performed with our HMM-based speech recognition system [3]. The gender independent, decision tree clustered acoustic models consisted of continuous Gaussian mixture densities. Evaluation tests were carried out with a single pass integrated trigram decoder based on word-internal triphones.

2.2. Description of quality measures

Our approach to detecting transcription errors (i.e. wrongly transcribed words, missing words) and incorrect segment boundaries is based on a forced Viterbi alignment on the training data and the evaluation of different transcription quality measures. First, low resolution acoustic models (2000 tied states, 60k densities) were trained on 46 hours of manually cleaned Hub-4 training data. The alignment was then carried out with our speech recognition system.

We investigated six criteria to detect erroneous training scripts. Segments were classified according to

(1) whether or not the optimal path in the DP time alignment did reach the terminal HMM state,

(2) the width of the beam required for the alignment,

(3) the acoustic sentence score, normalized to the number of time frames

(4) the normalized acoustic word score, and

(5) the duration of each word in the segment.

(6) In addition, adjacent segments were joined to compare the location of the boundary obtained by forced alignment with the boundary given in the training script.

Segments with bad quality according to these measures where sorted in order to inspect or reject the worst ones first.

2.3. Efficiency of the error detection criteria

Whereas the criteria (1), (4), and (6) proved to be useful in detecting script errors, there was only little correlation between segment quality and the criteria (2), (3), and (5).

Segment-wise criteria, (1)–(3): Measure (1) mainly detected major errors in training scripts like whole untranscribed sentences or incorrect segment boundaries. Measures (2) and (3) were highly speaker- and focus condition dependent and therefore of little use in detecting script errors.
Mistranscribed single words were not detected by any of these segment-wise criteria due to the usually long training segments. The sentence score of a given segment is the normalized sum of word scores. Hence, the poor acoustic score of one wrongly transcribed word may be masked by the scores from the other words. Equally, the DP algorithm may reach the terminal HMM state even if the

Viterbi path is not adequate at the beginning or middle of the segment.

Word-wise criteria, (4) and (5): Criterion (4) indicated missing or wrongly transcribed single words, but also utterances with strong background noise or overlapping speech. On the contrary, measure (5) gave only little evidence of script errors as the duration of words is basically speaker- and context dependent. Words with significantly shorter duration than their average were rarely observed, which could have been caused by the minimum word length constraint of our HMM models.

Finally, the **across-segment criterion (6)** indicated wrong segment boundaries as well as major transcription errors like measure (1).

As (2), (3), and (5) showed poor efficiency in detecting transcription errors, they were excluded from further analysis.

2.4. Application of the criteria

Starting from a forced Viterbi alignment we calculated the difference Δ between the final HMM state and the terminal state according to the script for each segment. Likewise, we calculated the normalized acoustic score s_w for each word in the segment and the time difference Δt between the segment boundaries of adjacent segments according to (6). We then computed the mean score $\bar{s}_w$, variance σ_w, and the number of observations N_w for each word as well as the overall mean $\bar{s}_{all}$ and variance σ_{all} of all word scores. A segment was considered to have potential script errors if

(1) $\Delta > 10$,

(4) $N_w > 10 : (s_w - \bar{s}_w)/\bar{s}_w > 3\,\sigma_w$
$N_w \leq 10 : (s_w - \bar{s}_{all})/\bar{s}_{all} > 3\,\sigma_{all}$,
and / or

(6) $\Delta t > 500$ ms.

That is, if a word did not occur frequently enough ($N_w \leq 10$) in the training corpus we used the overall mean word score $\bar{s}_{all}$ and variance σ_{all} as fallback values.

The deviations were considered to be significant only if they exceeded a certain value. The thresholds were chosen in such a way that about one third of the corpus was marked. This was the order of suspicious segments reported by other groups.

3. RESULTS

3.1. Segment classification statistics

We applied our method to 76 hours of speech data in 15 389 segments from the 1996/97 Hub-4 training corpus. We marked 35% (5 429 segments, 28h) as possibly erroneous. Most segments (72%) were tagged because of bad acoustic word scores (4). Criteria (6) and (1) supplied 13% and 7% of the bad segments, respectively. The remaining 8% were classified as bad according to two or all three criteria.

The marked segments were manually corrected afterwards. During the correction process we estimated that the rate of false alarms was in the order of 25%, which means that most segments labelled as 'bad' actually contained wrong transcriptions or segment boundaries. After correcting, 75 hours of training material remained; only one hour worth of data was considered to be too bad for training because of overlapping or unclear speech.

In order to investigate the number of errors that remained undetected we first examined a sample of 25% of the segments from CD 4 which were not marked. From these segments, 11% contained errors which were not detected by our method.

Additionally we analysed another 3.5 hour subset (CD 1) of the training corpus in more detail. All segments of this subset were manually corrected and scored according to four categories: minor, medium, and major script error, and too bad for training. When evaluating the performance of our quality measures we focused on the last three categories. Segments falling into the category 'minor' had errors like untranscribed noises or errors affecting only single phonemes like *get* ↔ *got*.

From 1 352 segments in this second subset our method automatically marked 286 (21%).

Table 2 shows detailed results for the different quality labels.

error type	# segments	# segments automatically detected	
minor	192	62	32%
medium	72	30	42%
major	20	12	60%
too bad	209	48	23%

Table 2: Statistics of automatic error detection on CD 1

These results do not confirm the impression we had when correcting the automatically marked segments. They also contradict the results obtained from examining the subset of CD 4. On CD 1 the rate of false alarms was 52%, significantly higher than the average. On the other hand, the percentage of automatically tagged segments (21%) was below the average of the whole corpus (35%).
A possible explanation could be that the data on CD 1 are of especially bad quality for some reason. The 48 segments automatically labelled as 'too bad' are 20.6% of this category of the whole 1996/97 training corpus (233 'too bad' segments), although CD 1 makes up less than 5% of the corpus. It might have been better not to compute mean and variance word scores for the whole corpus but rather per CD or even per show or speaker.

3.2. Effects on the word error rate

The transcription verification approach presented here was further verified by evaluation tests with three different training data sets:

- the complete 1996/97 Hub-4 training corpus which amounts to about 76 hours,
- the manually verified 46 hour subset, in which all incorrect segments were rejected and only a few obvious errors were corrected, and
- the 75 hour subset, where an overall of 22 hours of erroneous segments were automatically detected and manually corrected thereafter.

While our preliminary Hub-4 system performed clearly better when trained with clean but less data (table 1), we observed a different behaviour of the system that was optimized for this recognition task (table 3). These results will be discussed in the next section.

1996/97 Hub-4 training corpus	size	WER
manually corrected	46h	33.6%
complete corpus	76h	32.8%
automatically verified + manually corrected	75h	32.5%

Table 3: Recognition results on Hub-4 '96 eval. set, obtained with our LVCSR system *optimized* for the Hub-4 task using different training corpora.

4. DISCUSSION

The method proposed in this paper is able to detect major transcription errors. In about 50% of all cases the criteria even marked the position of the error correctly, at least within the range of a few words. We expect an improved performance by adjusting the thresholds for segment classification, which have not been optimized so far. Furthermore we intend to combine the quality criteria described here with confidence measures, which have been shown to reduce tagging error rates on different corpora [4].

As seen from our optimized Hub-4 system, the question of quality vs. quantity of training data is not easy to answer. It seems that our acoustic models are more robust on the difficult Hub-4 '96 eval. test set when trained on more but unclean data. It will be subject of further analysis to find out how many errors need to be detected (and corrected) in order to achieve the best recognition performance.

With further increase of training corpora size in future, manual correction will become infeasible and thus, increase the importance of reliable automatic error detection methods. Then the goal will be to accept or reject suspicious segments or even parts of them rather than manually correcting the scripts.

Our method will also be tested and optimized on automatically transcribed training corpora like the TDT-2 corpus of 800 hours speech data, which will soon be released. The transcriptions of such corpora will have a significantly higher error rate than manually transcribed ones. In addition, the error types may differ from what has been observed so far. Both will affect the performance of LVCSR systems and our verification approach. Thus, the automatic transcription verification will remain a challenging task in future.

References

1. J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *Proc. DARPA Speech Recognition Workshop*, Lansdowne, VA, Feb. 1998.
2. P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A.Tuerk, E.W.D. Whittaker, S.J. Young , "The 1997 HTK Broadcast News Transcription System", *Proc. DARPA Speech Recognition Workshop*, Lansdowne, VA, Feb. 1998.
3. H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: "The RWTH Large Vocabulary Continuous Speech Recognition System", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, pp. 853-856, May 1998.
4. F. Wessel, K. Macherey, R. Schlüter: "Using Word Probabilities as Confidence Measures", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, pp. 225-228, May 1998.