Using Untranscribed User Utterances for Improving Language Models based on Confidence Scoring

Mikio Nakano

NTT Communication Science Laboratories
NTT Corporation
3-1 Morinosato-Wakamiya
Atsugi, Kanagawa 243-0198, Japan
nakano@atom.brl.ntt.co.jp

Timothy J. Hazen

Spoken Language Systems Group
MIT Laboratory for Computer Science
200 Technology Square
Cambridge, MA 02139, USA
hazen@sls.lcs.mit.edu

Abstract

This paper presents a method for reducing the effort of transcribing user utterances to develop language models for conversational speech recognition when a small number of transcribed and a large number of untranscribed utterances are available. The recognition hypotheses for untranscribed utterances are classified according to their confidence scores such that hypotheses with high confidence are used to enhance language model training. The utterances that receive low confidence can be scheduled to be manually transcribed first to improve the language model. The results of experiments using automatic transcription of the untranscribed user utterances show the proposed methods are effective in achieving improvements in recognition accuracy while reducing the effort required from manual transcription.

1. Introduction

The recent advancement of speech and language technologies has made it possible to deploy speech interfaces for use by the general public. Most of the commercially available speech interfaces such as voice-portal services adopt directed-dialogue strategies, in which the system tightly controls the dialogue by asking questions which constrain the user to answer with short phrases. On the other hand, conversational systems that can understand less restricted user utterances, which possibly consist of dozens of words, are currently under research, and their performance has significantly improved over the past several years [1, 2].

One manually-intensive task in developing these kinds of systems is the transcription of user utterances collected by the system. These transcriptions can be very useful for improving the statistical language models of the speech recognizer. Transcribing utterances requires extra effort when dealing with languages such as Japanese in which there are no standard word boundaries and writings. Transcriptions need to be manually segmented into consistent words or application-dependent tools for segmenting transcriptions need to be developed. This is why most of the speech interface developers choose writing recognition grammars by hand to avoid this effort.

This work was done as a part of a collaboration project between MIT and NTT. Mikio Nakano participated in this research at MIT as a Visiting Scientist. This work was also supported by DARPA under contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center.

This paper presents a method for reducing the amount of effort for transcribing user utterances for training language models in conversational systems. Our approach assumes that a small number of transcribed utterances and a large number of untranscribed utterances in the working domain are available. This is a realistic situation that occurs soon after a prototype system has been deployed.

The method is a combination of two methods. The first improves the initial language model, trained only with the transcribed data, by utilizing automatically derived transcriptions of the untranscribed data. These automatically derived transcriptions are augmented with recognition confidence scores, which allows poorly recognized utterances to be removed before the language model is trained. In this paper, we call this process *unsupervised training*. This idea is very similar to Gretter and Riccardi's method [3], which adapts the language model trained from the corpus in one domain to another similar target domain using untranscribed utterances from the target domain. The second technique determines which utterances should be transcribed first, based on the confidence scores, in order to build better language models, when the amount of effort available for transcription is limited. This can be considered to be active learning, which has been applied to training language and acoustic models for speech recognition by Hakkani-Tür et al. [4]. Our method combines these two techniques by applying the first technique to the utterances which are not selected to be manually transcribed in active learning.

This paper also shows the results of the experiments in two domains in different languages, i.e., a flight travel planning domain in English [2] and a weather information domain in Japanese [5]. These results show that the proposed methods are effective in reducing the effort of transcription.

2. Approach

2.1. Confidence Scoring

In our approach, recognition confidence scores are required in order to determine which automatically transcribed utterances should be used when training the language model. The recognition confidence scoring technique we employ is described in detail in [6]. Although confidence scores for both utterances and words can be obtained, only those for words are used in this paper. The confidence scoring technique produces zero-centered log-likelihood ratios, where positive scores indicate a high likelihood of a hypothesized word being correct, while negative scores indicate a high likelihood of a hypothesized word be-

ing incorrect. Note that a process which is called *confidence model training* is required to optimize parameters for computing confidence scores so that the error rate for the confidence classification on *unseen* data is minimized.

2.2. Unsupervised Training: Language Model Improvement using Automatic Transcriptions

For improving the language model, it is desirable to utilize untranscribed utterances that have only a small number of hypothesized words with low confidence. To achieve this, we only use utterances that satisfy the condition that the ratio of the hypothesized words whose confidence scores are lower than t is not greater than r. Word hypotheses with low confidence are replaced by the marker '⟨unknown⟩', which is handled as an out-of-vocabulary word during language model training.

An issue that must be resolved is how to determine the thresholds t and r. If t is high, the ratio of wrong hypotheses will be low, causing the number of utterances accepted for training to be limited. Because the number of accepted hypotheses may be small, they might not be effective for improving the language model. If r is high, the number of hypotheses accepted for training increases, but these hypotheses may include many '⟨unknown⟩' markers, which might be harmful to the n-gram probability estimates.

To determine the optimal values of t and r, a fraction of utterances are jackknifed from the transcribed set for testing. The language model and the confidence model are trained with the remaining transcribed utterances. The thresholds are selected to minimize the word error rate on the jackknifed test set.

Below are the steps we employ for utilizing untranscribed utterances to improve the language model. We assume that none of the utterances are used to train the recognizer's acoustic models, i.e., a pre-existing set of generic acoustic models trained from other data sources is available.

1. Split the transcribed user utterances into three sets, i.e. the *initial training set* (INIT), the *confidence model development set* (CMD), and the *language model development set* (LMD).
2. Build the initial recognizer with the language model trained with the transcriptions of the utterances in the INIT set.
3. Train the confidence model using the word hypotheses generated from the initial recognizer when tested on the CMD set.
4. Rebuild the recognizer with the language model trained with the transcriptions of the utterances in both the INIT and CMD sets.
5. Recognize the set of untranscribed utterances (the TRAIN set) and compute confidence scores for the hypothesized words using the recognizer and the confidence model.
6. Determine two thresholds t and r by executing the following steps for a variety of values of t and r and selecting the values that give the best result.
   (a) Given thresholds t and r, split the untranscribed utterances into the following two classes:
      A Utterances in which the ratio of the hypothesized words whose confidence scores are lower than t is not greater than r. This subset of the TRAIN set is called accepted(TRAIN) hereafter.
      B Other utterances.

Table 1: *The number of utterances in each set.*

Set	MERCURY	MOKUSEI
INIT (CMD)	866	803
LMD	791	802
TRAIN	16,855	6,160
Test set	1,677	2,500

   (b) Let auto(accepted(TRAIN)) be the collection of automatically derived transcriptions for the accepted(TRAIN) utterances for which the words whose confidence score is lower than t are replaced by '⟨unknown⟩'.
   (c) Train the language model with the manual(INIT), manual(CMD) and auto(accepted(TRAIN)) sets, where manual(⟨SET⟩) is the collection of manual transcriptions of utterances in the set ⟨SET⟩.
   (d) Recognize utterances in the LMD set and compute the recognition accuracy.
7. Rebuild the recognizer with the language model trained with the manual(INIT + CMD + LMD) and auto(accepted(TRAIN)) sets, where the utterances in accepted(TRAIN) are chosen based on the optimal thresholds found for t and r on the LMD set.

2.3. Active Learning: Selecting Utterances to Transcribe

Since the above method is based on the recognition hypotheses, it has the problem that utterances in which the initial language model gives low probability are difficult to recognize. As a result, the automatically derived transcriptions for these utterances are likely to be filtered out due to poor confidence, and hence not reflected in the language model training. Therefore, if possible, it would be most effective to manually transcribe the specific utterances whose inclusion in the training set would most improve the language model, rather than randomly selecting utterances for manual transcription. This process can be considered as a kind of *active learning*.

Rather than using just active learning as Hakkani-Tür et al. [4] do, we consider using it together with the unsupervised training described above. One possibility is to manually transcribe the utterances that are not used for the unsupervised training (i.e., TRAIN - accepted(TRAIN)), because hypotheses for accepted utterances can be effectively used for language model training. We discovered that the rejected utterances included many short utterances which might not be effective for trigram learning. In addition, the number of rejected utterances may be large, and we need an additional criterion to choose utterances from this class.

We therefore assume that transcriptions of utterances whose recognition hypotheses include a larger number of low-confidence word hypotheses are more effective for improving the language model than those of other utterances. If the number of the low-confidence word hypotheses is large, the average length of the utterance will also be large.

Our method is described formally as follows.

1. Initially set m to be the maximum number of words within the automatic transcriptions of the TRAIN set.

2. Manually transcribe additional utterances from the TRAIN set and use them for language model training if the number of word hypotheses whose confidence scores are less than t is greater than or equal to the threshold m. This set of utterances is denoted by select(TRAIN).
3. Recompute the confidence model and confidence scores for the hypothesized words for the TRAIN set.
4. Train the language model with:

   manual(INIT + LMD + select(TRAIN))
   + auto(accepted(TRAIN - select(TRAIN))).

5. If additional transcription effort is available, decrement m by 1 and go back to the step 2.

3. Experiments

To show the effectiveness of the proposed method, we conducted experiments on utterances in two domains, the MERCURY air-travel system (for English) [2] and the MOKUSEI weather information system (for Japanese) [5]. Both systems use SUMMIT, a segment-based speech recognizer [7]. For these experiments, the acoustic models are trained on data that do not include any utterances used in these experiments. The acoustic models of the MERCURY recognizer were trained from nearly 115,000 utterances, of which over 93% are from the JUPITER weather information domain [1], and that of the MOKUSEI recognizer was trained from about 3,000 expert user utterances for MOKUSEI and about 2,000 read utterances. This simulates the rapid development scenario where acoustic models are borrowed from pre-existing systems. Note that, because these models do not make use of the available domain-dependent data for acoustic model training, the results reported in this paper are worse than the performance of the actual recognizers used in the deployed versions of these systems. Both recognizers use class trigram language models. The language model of the MERCURY recognizer has 1,524 vocabulary entries in 51 classes and that of the MOKUSEI recognizer has 1,262 vocabulary entries in 57 classes.

For each of these two domains, we split the naive user utterances into four sets, INIT, LMD, TRAIN, and the test set. Table 1 shows the number of utterances in each set. In this experiment, we used the same set for INIT and CMD, in order to increase the number of utterances in these sets without increasing the number of transcribed utterances. In this case CMD is not unseen data in terms of the language model for the recognizer used for the confidence model training. However, because the confidence model relies primarily on the acoustic model scores, the effect of the language model may not be large when training the confidence model. We have not observed any ill-effects from this decision in our experiments.

To find optimal threshold values for t and r, we examined the word error rates over all data in the LMD test set using various combinations of values. The results are depicted in Fig. 1. We selected the threshold pair that gave the minimum word error rate, i.e., $t = -1$ and $r = 0.25$ for the MERCURY domain and $t = 1$ and $r = 0.25$ for the MOKUSEI domain. At these thresholds, roughly 37% of the automatically transcribed utterances are rejected in the MERCURY domain while 46% are rejected in the MOKUSEI domain. We experimentally confirmed that small changes in these thresholds do not significantly affect the later evaluations.

After selecting the optimal values of t and r we rebuilt the recognizer with the language model trained with the manual(INIT+LMD) and auto(accepted(TRAIN)) sets, and evaluated its performance. For comparison, we also examined how the language model is improved when the amount of transcribed utterances added to the initial training set is increased. The results are shown in Table 2. For the MERCURY domain, the performances of the resulting language model using 16,855 automatically transcribed utterances are comparable to those of the language model trained with manual transcriptions of 1,600 utterances in TRAIN (in addition to manual(INIT+LMD)). This means the proposed method is effective in reducing the effort of transcription, but at the cost of requiring nearly ten times more data. When only half of the automatically transcribed utterances in the TRAIN set are used, the reduction in the word error rate is smaller. This suggests that a suitably large enough TRAIN set is crucial.

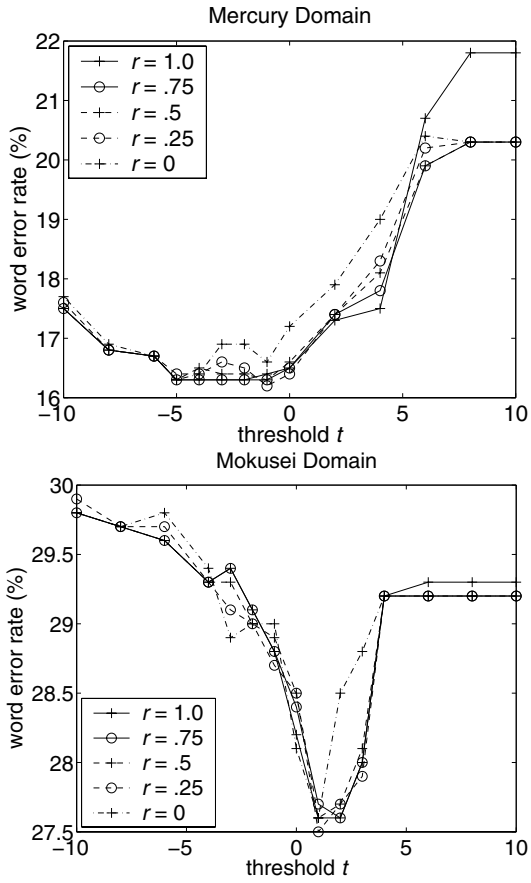

Figure 1: *Change in speech recognition performance over the LMD set depending on changes in thresholds. See text for definitions of r and t.*

Next, we conducted another experiment to show the effectiveness of our active learning method. In this experiment, we only used MERCURY data for the following reason. In the MOKUSEI domain, there is not much difference between the performance of the language model trained with the manual transcriptions of all of the TRAIN (WER of 27.0%) and that of the language model trained with the automatic transcription selected with our method (WER of 27.4%), thus, additional manual transcriptions are not very helpful for improving the language model.

In this experiment, we omit step 3 to simplify the procedure.

Table 2: *Improvement in speech recognition performance with the confidence-scoring-based hypothesis utilization and its comparison with the results obtained by adding manual transcriptions.*

training data	word error rate (%)	
	MERCURY	MOKUSEI
manual(INIT+LMD)	22.2	28.9
+manual(800 in TRAIN)	21.3	27.6
+manual(1,600 in TRAIN)	20.3	27.6
+manual(3,200 in TRAIN)	19.6	27.2
+manual(6,400 in TRAIN)	18.8	n.a.
+manual(12,800 in TRAIN)	18.0	n.a.
+manual(TRAIN)	17.7	27.0
+auto(accepted(half of TRAIN))	21.2	27.5
+auto(accepted(TRAIN))	20.4	27.4

Instead, we fixed the thresholds t and r respectively to -1 and 0.25 as determined in the experiment described above. Then we examined how the language model gets improved when we increase the amount of manual transcriptions with decrementing the threshold m, by investigating the relationship between the size of select(TRAIN) and the resulting speech recognition performance. The results are shown in Fig. 2 as the select+auto line. Since the amount of effort to transcribe an utterance increases with the utterance length, we normalized for this by taking the average number of words in the utterances into consideration. This is shown in the figure as the select+auto (normalized) line. For comparison, we examined the recognition performances when additional manually transcribed utterances are selected randomly. If this set of randomly selected utterances is called random(TRAIN), the full set of training utterances for this condition can be expressed as:

manual(INIT + LMD + random(TRAIN))
+ auto(accepted(TRAIN - random(TRAIN))).

The result for this training set is shown in the figure as the random+auto line. We also examined the result of adding only hand transcribed utterances. This is shown as the random line in the figure and its set can be expressed as:

manual(INIT + LMD + random(TRAIN)).

In this graph, both of the select+auto and select+auto (normalized) lines are, for the most part, beneath the random+auto line. This indicates that the selection of utterances for manual transcription based on confidence scoring makes it possible to achieve the same recognition performance with fewer additional manual transcriptions, and thus it is effective in reducing the effort of transcription.

4. Concluding Remarks

This paper presented methods for improving the language model for the speech recognizer in conversational systems by using untranscribed user utterances. The effectiveness of the methods has been shown by experiments in two domains. Among future work is exploring how the results change when the sizes of the transcribed and untranscribed data sets are varied. We are hoping to establish some criteria for determining the appropriate set sizes through experiments and theoretical considerations.

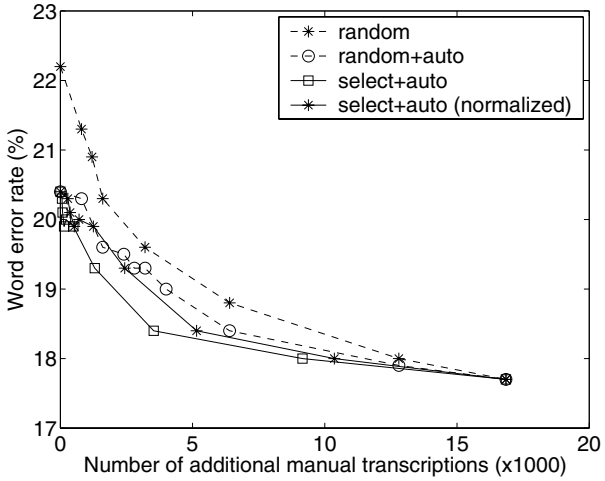

Figure 2: *The relationship between speech recognition performance on the test set and the amount of manual transcriptions besides the INIT and LMD sets.*

5. Acknowledgements

The authors would like to thank everyone who contributed to building MERCURY and MOKUSEI and collecting data using them, including Scott Cyphers, Jim Glass, Lee Hetherington, Yasuhiro Minami, Joe Polifroni, Stephanie Seneff, and Victor Zue.

6. References

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "Jupiter: A telephone-based conversational interface for weather information," *IEEE Trans. Speech and Audio Proc.*, 8(1):85–96, January 2000.

[2] S. Seneff, "Response planning and generation in the MERCURY flight reservation system," *Computer Speech and Language*, 16(3–4):283–312, 2002.

[3] R. Gretter and G. Riccardi, "On-line learning of language models with word error probability distributions," in *Proc. ICASSP*, 2001.

[4] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. ICASSP*, 2002.

[5] M. Nakano, Y. Minami, S. Seneff, T. J. Hazen, D. S. Cyphers, J. Glass, J. Polifroni, and V. Zue, "Mokusei: A telephone-based Japanese conversational system in the weather domain," in *Proc. Eurospeech*, 2001, pp. 1331–1334.

[6] T. J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language*, 16:49–67, January 2002.

[7] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," in *Proc. ICSLP*, 1996, pp. 2277–2280.