LANGUAGE MODEL ADAPTATION FOR AUTOMATIC CALL TRANSCRIPTION

Ali Haznedaroglu[1,2] and Levent M. Arslan[1,2]

[1] Sestek, Istanbul, TURKEY
[2] Electrical and Electronics Engineering Department, Bogazici University, Istanbul, TURKEY
{ali.haznedaroglu, levent.arslan}@sestek.com

ABSTRACT

This paper presents a method of language model adaptation for call-center conversations using automatic speech recognition (ASR) transcripts and their confidence scores. The goal is to select the optimal adaptation set by estimating the recognition errors and minimizing the adaptation language model (LM) perplexity. ASR transcripts are ranked with respect to their confidence scores and adaptation data selection is done iteratively by filtering the most reliable transcript set that minimizes the LM perplexity. Model adaptation is then carried out by interpolating the selected adaptation LM with the baseline in-domain LM. We have evaluated our approach on agent speech of real call-center conversations and experiments show that 4% relative word error rate reduction is achieved with the proposed approach.

Index Terms— language modeling, large vocabulary continuous speech recognition, language model adaptation, speech analytics

1. INTRODUCTION

Building state-of-the-art speech recognition systems is a highly data-driven process, which requires large amounts of in-domain training data to be transcribed by human supervision. Required amount of manual data transcription further increases in the case of spontaneous speech recognition, as less digitized data is available and the recognition accuracies significantly depend on the training data size [1]. However, manual data transcription is a labor intensive process and is not feasible in many cases because of cost or time restrictions. In such cases, unsupervised learning or adaptation methods can be utilized to minimize the human supervision effort and cost, while improving the system performance.

Language model (LM) adaptation is used in cases where the existing in-domain training data is insufficient to reliably model the n-gram statistics, or even where no prior in-domain data is available [2]. Both cases are investigated in [3] on an unsupervised domain adaptation task, where automatic speech recognition (ASR) transcripts are used as the unsupervised training set. This work also investigated and compared two LM adaptation approaches, namely model interpolation and count mixing. In [4], unsupervised LMs are built from scratch, by investigating the word-error probability distributions of the ASR outputs, and filtering the candidate transcripts based on these confidence measures (CM). In [5], unsupervised LM adaptation is employed for automatic transcription of meeting recordings where no in-domain training data is available. A generic out-of-domain LM is used to recognize adaptation data to estimate an in-domain LM from the recognition transcripts, which is then interpolated with the generic out-of-domain to obtain the adapted LM. A supervised LM is also trained from manually transcribed in-domain data to evaluate the effects of ASR errors introduced during the unsupervised adaptation, and the results showed that final accuracy improvement is doubled when supervised adaptation is used instead of unsupervised adaptation. Unsupervised adaptation and active learning paradigms are used for call classification in [6], where an iterative process is employed to improve the recognition accuracy. Unsupervised LM training mainly follows the same procedure in [5], and for active learning, CMs are used to identify candidate utterances that are needed to be manually transcribed. The impact of manually transcribed data amount on the recognition performance is also measured in this work. Active and supervised learning methodologies for speech recognition are further investigated in [7], [8]. Unsupervised LM adaptation for broadcast news recognition is investigated in [9], where ASR transcripts are used as queries for information retrieval based adaptation data selection from a general corpus. Direct likelihood maximization selection (DLMS) from the ASR transcripts is used for LM adaptation in [10], which aims to minimize the effects of recognition errors on the adaptation data. If the domain diversity of the training data is high, topic models can be used to adapt the LM domain to the test set domain. Latent Dirichlet allocation (LDA) is used in [11] to assign topic weights to training set n-grams, which are then multiplied by global n-gram counts to obtain the topic adaptation LM. Lecture recognition is another area where topic-diversity is high and lecture-specific LM adaptation is needed to optimize the ASR systems. In [12], lecture related documents are automatically retrieved from the world-wide-web (WWW), and lecture specific LM and vocabulary adaptation are done using this collected corpus. A detailed work on data collection and language model adaptation from WWW resources can be found on [13].

In this work, we focus on LM adaptation for Turkish call-center conversation recognition. Call-center conversation recognition is still a challenging task even for the state-of-the art speech recognition systems. Other than the acoustical difficulties introduced by the telephone channel effects or environmental noises, reliable LM estimation is challenging because of

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

spontaneous characteristics of the conversations. These conversations may include hesitations, repetitions, and partial words, so they may not be linguistically well-structured [1]. Free word order characteristic of sentence formation in Turkish also adds further difficulties in our case, as the average word branching factor increases which also results in an increase in the language model perplexity [14]. In our scenario, manually transcribed in-domain data is available, and unsupervised LM adaptation on this data is investigated in order to obtain further recognition improvements. Our method involves constructing an adaptation set from ASR outputs of real call-center conversations, and in order to minimize the effects of recognition errors on the adapted LM estimation, we iteratively filter the adaptation set with respect to the confidence scores obtained from the ASR output lattice. This iterative process allows us to choose an optimal adaptation set which minimizes the LM perplexity. Adaptation data selection using word or utterance based confidence scores is investigated in [4], [7]. This work extends the word or utterance level confidence score filtering to a whole conversation level, and uses a perplexity minimization approach to iteratively select the adaptation set. This work can also be regarded as an initial effort of LM adaptation for Turkish conversational speech recognition, and shows how adaptive learning strategies can improve recognition accuracies of such systems over time, without any human involvement.

2. APPROACH

In this section, we review LM adaptation and ASR confidence score estimation concepts. These concepts act as building blocks of our adaptation approach which is also described in this section.

2.1. Language Model Adaptation

ASR systems can be seen as Bayesian classifiers that choose the most likely word sequence, given an acoustic observation, and prior acoustic and linguistic knowledge. This problem can be mathematically formulated as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(X|W)\, P(W|\Lambda)$$

Here X is the observed acoustic sequence and $\hat{W}$ is the most likely word sequence at the output of the recognizer. In this equation, P(X|W) models the acoustic observation probability, and is termed as the acoustic model. The other term, P(W|Λ) provides a priori probability of the word sequence given a linguistic knowledge, Λ. In practice, this term is specified by an n-gram language model. Language model training requires large amount of in-domain data for reliable computation of the n-gram statistics. In situations where in-domain training data is not available or not sufficient, language model adaptation attempts to obtain more reliable models for the target domain.

Model interpolation or n-gram count mixing can be used in LM adaptation. In this work, we use model interpolation, in which n-gram statistics of two models are interpolated to form a new model:

$$P_{adapt} = \gamma\, P_1 + (1\text{-}\gamma)\, P_2$$

where P_j's are the n-gram probabilities of the source models, and P_{adapt} represents the n-gram probabilities of the adapted model. Interpolation weight γ, can be optimized through Expectation-Maximization (EM) algorithm on a development test set.

2.2. Confidence Measures

When ASR transcripts are used as adaptation data, recognition errors introduce noise on the n-gram counts so that adapted LM cannot be estimated reliably. In such cases, confidence measures can be used as error predictors which are used in adaptation data selection by filtering out error prone transcripts. Different CM approaches are deeply investigated in [15]. In this work, we calculate the posterior probability from the word lattice output of the recognizer, and use it as a CM. In [15], this confidence estimation method is observed to achieve better performance than the other methods.

In order to verify the error estimation performance of the CM used, we run a separate test, in which confidence scores are compared with respect to the recognition accuracies. A similar ASR setup to [16] is used in this test, and the test set contains 1800 call center agent utterances. Although their domains are identical, this test set is different from the sets we use in further LM adaptation experiments. The test is done as follows. First, the utterances are recognized, and ranked with respect to their confidence scores. Then, this ranked list is divided into 10 bins having equal number of utterances. The first bin contains the utterances with the highest confidence scores, and the confidence scores decrease as the bin number increases. Then, average word error rate (WER) for each bin is calculated and the results are shown in Figure 1, which verifies the error estimation performance of the employed CM.

This work focuses on the agent speech of the call center conversations, and a confidence score is calculated for every call. First, the call conversation is partitioned into agent and customer speech. Then agent speech part is separated into its voice and silence segments using a voice-activity-detection module. Only the voiced parts are recognized, and their duration-normalized posterior probabilities are accumulated in order to obtain the final confidence score for each call. Calculating the entire conversation confidence scores by duration normalized posterior probability accumulation gives the best results in our case.

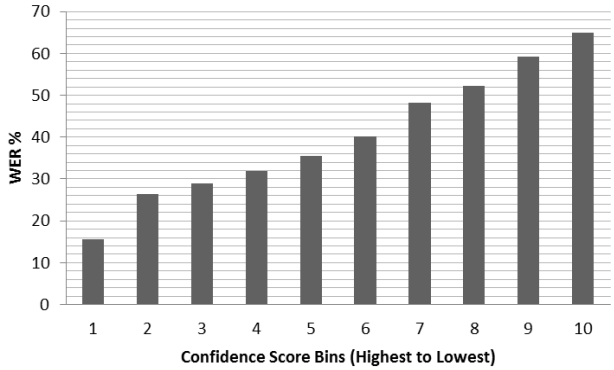

Figure 1. WER vs. ranked confidence scores.

2.3. Adaptation Approach

The steps involved in our adaptation approach are as follows:

1. A baseline language model Θ_B is built from manually transcribed in-domain data S_B.
2. Adaptation data is recognized using the baseline model Θ_B, which outputs the adaptation transcript set $\hat{S}_A$ and a list of corresponding confidence scores. Each item in $\hat{S}_A$ represents the transcription of the agent speech of the entire call, and its confidence score.
3. Transcripts in $\hat{S}_A$ are ranked with respect to their confidence scores.
4. Transcripts in $\hat{S}_A$ are iteratively filtered with respect to their confidence scores. At each iteration, top *i* transcripts with the highest confidence scores are selected as the training data, from which a candidate ASR transcription model Θ_i is built. Perplexity of Θ_i on the development set S_D is measured at every iteration, and the final transcription model is chosen as the one that yields the minimum perplexity.
5. Interpolation weights are chosen using EM algorithm on the development set S_D, and baseline model Θ_B and the selected transcription model Θ_i are interpolated to obtain the adapted model Θ_A.
6. Baseline model Θ_B and adapted model Θ_A are tested on the test set S_T, and their recognition accuracies are compared.

3. EXPERIMENTS

3.1. Call Center Conversation Data

In the baseline LM training 2000 call center recordings are manually transcribed and used. Total duration of this set is 150 hours. Adaptation data consists of 38000 hours of call-centers recordings, which are automatically recognized by an HMM based speech recognizer based on CMU Sphinx toolkit [17]. Development and test data consists of 3 and 9 hours of manually transcribed call center recordings respectively. Each data set is collected at different time intervals from the call center so they are non-overlapping. This work focuses only on the agent speech of the conversations, and all the manual and automatic data transcriptions are done on the agent speech only. Some statistics of the call recordings used in this work are given in Table 1.

Average call duration	270 seconds
Average call silence duration	119 seconds
Average agent speech duration	85 seconds
Average number of agent words per call	230

Table 1. Statistics of calls used in the experiments.

3.2. Recognition System

Acoustic model trainings and recognitions are done using the CMU Sphinx toolkit [17]. A speaker-independent acoustic model is trained by using approximately 1000 hours of call center conversations, recorded in different call centers. Only 150 hours of this data is used in baseline language model training, as its call center domain matches to that of the test set, so that the language model estimation is better. Acoustic training data also does not overlap with the adaptation, development and test data used in LM adaptation experiments. Feature vectors consist of 12 Mel-frequency cepstral coefficients and energy, together with their first and second order derivatives. Cepstral mean normalization is applied to the feature sets in order to minimize the transmission channel differences. Context-dependent triphone models with 6000 tied-states using 12 Gaussian mixtures are then trained for a Turkish phoneme set. One-best hypothesis of each utterance is outputted by the recognizer, together with the word lattice that is used in CM calculations.

LM training and adaptations are done using the IRSTLM toolkit [18]. Word-based, 3-gram LMs are built using Witten Bell smoothing in all cases. Acoustic model, dictionary, and recognition parameters are kept fixed in all of the recognition experiments. Recognition dictionary contains approximately 20k words that occur in the manually transcribed in-domain data. Out-of-vocabulary rate of the test set is 3%.

3.3. Results

The Adaptation set is recognized using the baseline LM, and the output transcripts are sorted with a descending order according to their confidence scores. Then iteratively, top *i* transcripts are selected and a candidate adaptation model is built. Figure 2 shows the perplexity scores of these candidate models on the development set as *i* increases. Minimum perplexity is obtained when 10k call transcripts with the highest confidence scores are used. This set is equivalent to approximately 750 hours of call-center conversations.

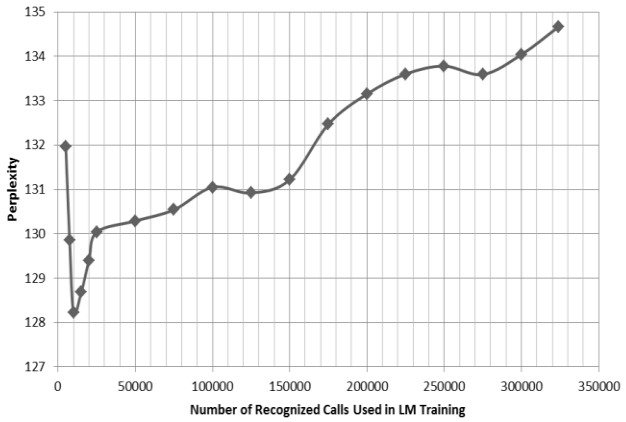

Figure 2. Perplexity with respect to confidence score based adaptation data selection

Iteratively selected adaptation LM is then interpolated with the baseline LM to obtain the final adapted LM, and Table 2 presents the perplexity reduction on the development set when the baseline LM is interpolated.

Language Model	*Perplexity*
Baseline (2k calls)	45.50
Selected adaptation set (10k calls)	128.23
Adapted (2k + 10k calls interpolated)	38.15

Table 2. Perplexities of the language models trained.

Recognition accuracies of the LMs on the test set are presented in Table 3. Language model adaptation reduced the WER from 28.72% to 27.56% when 10k call transcripts with the

highest confidence scores are used as the adaptation data. The relative WER reduction is 4%. This table also shows the recognition accuracy when the LM is built using the 10k call transcripts. In such a case, where only the recognition outputs are used to train the LM, there is an absolute 3.68% increase in the error rate.

In order to verify our iterative adaptation set perplexity minimization approach, we also interpolated our baseline LM with other candidate adaptation LMs which have relatively low perplexities on the development set. Recognition accuracy experiments on the test set confirm that best WER reduction is obtained when 10k call transcripts are used in LM interpolation, as shown in Figure 3.

Language Model	*WER*
Baseline (2k calls)	28.72%
Selected adaptation set (10k calls)	32.40%
Adapted (2k + 10k calls interpolated)	27.56%

Table 3. Recognition accuracies of the language models.

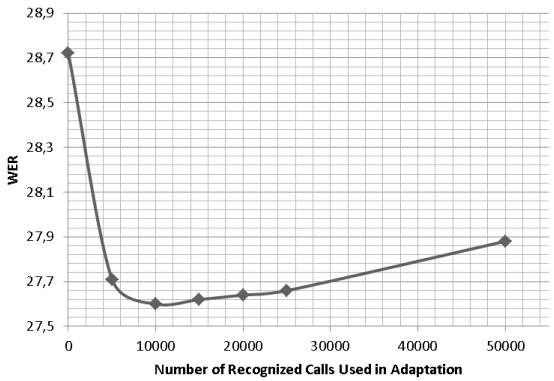

Figure 3. Recognition accuracies with respect to the number of adaptation transcripts used in LM interpolation.

4. CONCLUSIONS

In this work, we presented a language model adaptation approach, which uses recognition transcripts and their confidence scores to select the adaptation data by perplexity minimization criterion. Posterior probabilities, which were calculated from the recognition lattices, were chosen as the confidence measures and we have empirically shown that these CMs are good error estimators. Adaptation data selection was done iteratively by filtering the most reliable transcript set that minimizes the LM perplexity. Tests showed that top 10k call transcripts with the highest confidence scores minimizes the perplexity, and 4% relative word error rate reduction was achieved when this selected set is interpolated with the baseline LM.

One question that is not answered in this paper is, to what extent the recognition accuracy may improve when same amount of manually transcribed data is used in LM adaptation. Manual transcription of such amount of data is a long-lasting process, so this case may be investigated in future. Future work will include extending our approach also to the customer speech of the conversations. Customer speech may not be linguistically well-structured as agent speech, so it can be more challenging for the adaptation method to succeed on such higher perplexity conditions.

5. ACKNOWLEDGMENTS

This work is supported in part by TUBİTAK TEYDEB 1509 under the Electronic Doctor's Round Project, number 9090036**.**

6. REFERENCES

[1] Sadaoki Furui, et al. "Why is the recognition of spontaneous speech so hard?," in *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2005.

[2] Jerome R. Bellegarda, "Statistical language model adaptation: review and perspectives," in *Speech communication*, vol. 42.1, pp. 93-108**,** 2004.

[3] Michiel Bacchiani and Brian Roark. "Unsupervised language model adaptation," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, pp. 224-227, 2003.

[4] Roberto Gretter and Giuseppe Riccardi. "On-line learning of language models with word error probability distributions," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, pp. 557-560, 2001.

[5] Gokhan Tur and Andreas Stolcke. "Unsupervised Languagemodel Adaptation for Meeting Recognition," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 4, pp. 173-176, 2007.

[6] Dilek Hakkani-Tur, et al. "Unsupervised and active learning in automatic speech recognition for call classification," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, pp. 429-432, 2004.

[7] Giuseppe Riccardi and Dilek Z. Hakkani-Tür. "Active and unsupervised learning for automatic speech recognition," in *Proc. EUROSPEECH*, 2003.

[8] Giuseppe Riccardi and Dilek Hakkani-Tur. "Active learning: Theory and applications to automatic speech recognition," in *IEEE Trans. on Speech and Audio Processing*, vol. 13.4, pp. 504-511, 2005.

[9] Langzhou Chen, et al. "Unsupervised language model adaptation for broadcast news," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, vol. 1, pp. 220-223, 2003.

[10] Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa. "Unsupervised CV language model adaptation based on direct likelihood maximization sentence selection," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, pp. 5029-5032, 2012.

[11] MDHaidar, and Douglas O'Shaughnessy. "Topic n-gram count language model adaptation for speech recognition," *In Proc. IEEE Spoken Language Technology Workshop (SLT)*, pp. 165-169, 2012.

[12] Paul Maergner, Alex Waibel, and Ian Lane. "Unsupervised vocabulary selection for real-time speech recognition of lectures," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, pp. 4417-4420, 2012.

[13] Abhinav Sethy, et al. "An iterative relative entropy minimization-based data selection approach for n-gram model adaptation," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17.1pp. 13-23, 2009.

[14] Ebru Arısoy, Helin Dutağacı, and Levent M. Arslan. "A unified language model for large vocabulary continuous speech recognition of Turkish," in *Signal Processing,* vol. 86.10, pp. 2844-2862, 2006.

[15] Hui Jiang, "Confidence measures for speech recognition: A survey." in *Speech Communication*, vol. 45.4, pp. 455-470, 2005.

[16] Ali Haznedaroglu, Levent M. Arslan, Osman Buyuk and Mustafa Erden, "Turkish LVCSR system for call center conversations," in *Proc. IEEE Conference on Signal Processing and Communication Applications – SIU*, pp. 372-375, 2010.

[17] P. Placeway, et al. “The 1996 hub-4 sphinx-3 system,” in *DARPA Speech Recognition Workshop*, Chantilly, 1997.

[18] M. Federico, N. Bertoldi, M. Cettolo, IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models, Proceedings of Interspeech, Brisbane, Australia, 2008.