LANGUAGE MODEL PARAMETER ESTIMATION USING USER TRANSCRIPTIONS

Bo-June (Paul) Hsu and James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA, 02139 USA
{bohsu,glass}@csail.mit.edu

ABSTRACT

In limited data domains, many effective language modeling techniques construct models with parameters to be estimated on an in-domain development set. However, in some domains, no such data exist beyond the unlabeled test corpus. In this work, we explore the iterative use of the recognition hypotheses for unsupervised parameter estimation. We also evaluate the effectiveness of supervised adaptation using varying amounts of user-provided transcripts of utterances selected via multiple strategies. While unsupervised adaptation obtains 80% of the potential error reductions, it is outperformed by using only 300 words of user transcription. By transcribing the lowest confidence utterances first, we further obtain an effective word error rate reduction of 0.6%.

Index Terms— speech recognition, language modeling, adaptation

1. INTRODUCTION

In an ideal world, language model parameters would be learned on training data that exactly matches testing conditions. In practice however, many potentially valuable applications for speech processing have only limited amounts of matched training data. In extreme cases, no such data exists beyond the unlabeled test data itself. One example of this latter situation are recordings of academic lectures. With the increasing availability of these kinds of video materials, accurate transcripts are needed to improve the search, navigation, summarization, and even translation of the content. Unfortunately manual transcription of the lecture audio is time consuming, expensive, and error-prone (e.g., *Markov* → *mark of*, *Fourier* → *for your*), so there is significant interest in using speech recognition technology to provide automatic transcriptions.

Compared with other types of audio data, lecture speech often exhibits a high degree of spontaneity and focuses on narrow topics with special terminologies [1]. While we may have existing transcripts from general lectures or written text on the precise topic, data that matches both the topic and style of the target lecture and speaker rarely exist. In off-line lecture transcription scenarios, we can perform multiple recognition passes over the lecture. Thus, the recognition hypotheses from previous iterations may be used as data approximating the target domain. In addition, since users are often motivated to obtain the most accurate transcription possible, they may be willing to transcribe a subset of the lecture utterances to aid the recognition. The resulting user transcriptions not only serve as in-domain development data, but also eliminate the effect of any recognition errors among these utterances.

Past language modeling research with sparse training data has investigated various adaptation and interpolation techniques that make use of partially matched corpora [2, 3, 4]. However, they generally assume the existence of an independent in-domain development set for parameter tuning. Several researchers have explored using the recognition hypotheses for language model adaptation. Typically, the hypotheses are used to build a component LM to be combined with the baseline LM via linear interpolation [5, 6] or count merging [7]. However, in certain settings, minimum discrimination information adaptation has been shown to yield slightly better results [8, 9]. Since the recognition hypotheses change through adaptation, the process can be repeated iteratively for potentially further improvements [7, 8].

In existing work, model adaptation parameters are often chosen arbitrarily. With only one parameter that specifies the weight of the adaptation data with respect to the baseline LM, tuning the perplexity on an accurate development set is not always critical [6]. However, when the baseline model is itself interpolated and contains multiple parameters, having an error-free development set to optimize both the baseline and adaptation parameters becomes paramount, as the parameters are now more likely to fit the errors. Thus, in this work, we extend previous work by considering an interpolated LM baseline with more sophisticated modeling techniques.

To obtain accurate in-domain data for tuning, we propose selecting utterances from the target lecture for transcription. Although active learning techniques have been proposed for selecting training utterances [10] based on confidence scores, the selection criteria in this case need to balance between eliminating errors from the transcribed utterances and building a representative development set. In this study, we will measure the effect of such tradeoff using multiple utterance selection techniques.

Dataset	# Words	# Sents	# Docs
Textbook	131,280	6,762	271
Lectures	1,994,225	128,895	230
CS	87,527	3,611	10

Table 1. Summary of evaluation corpora.

In this paper, we will describe research that explores various supervised and unsupervised approaches to tune the model parameters, with a focus on language modeling for lecture transcription. We first evaluate the effectiveness of using the recognition hypotheses as development data. Next, we study the performance of using various amounts of user transcription for parameter tuning. As the system can select the order in which utterances are presented to the user for transcription, we compare the effectiveness of transcribing utterance by chronological order, random sampling, and lowest utterance confidence. We apply the above techniques to various LM interpolation and weighting schemes [3, 4]. With only 300 words of user transcription, we were able to outperform unsupervised adaptation using the recognition hypotheses. By transcribing the utterances in increasing confidence, we further reduced the effective word error rate (WER) by 0.6%.

2. EXPERIMENTS

2.1. Setup

In this work, we evaluate the WER of various trigram LMs trained using the recognition hypotheses and user transcriptions on a lecture transcription task [11]. The target data consists of 10 lectures from an introductory computer science course (CS). For training, we consider the course textbook with topic-specific vocabulary (Textbook) and numerous high-fidelity transcripts from a variety of general seminars and lectures (Lectures). Table 1 summarizes all the evaluation data.

To compute WER, we use a speaker-independent speech recognizer [12] with a large-margin discriminative hierarchical acoustic model [13]. The lectures are pre-segmented into utterances via forced alignment against the reference transcripts [14]. The MITLM toolkit [15] is used to tune the model parameters to minimize the development set perplexity and compute the WER of the tuned models via lattice rescoring. We evaluate each of the 10 target lectures in (CS) independently and present the averages of the WER results.

In the following sections, we evaluate the effectiveness of using the recognition hypotheses and user transcriptions for parameter estimation on a variety of LM estimation and interpolation schemes. As a baseline, we consider trigram models smoothed using fixed-parameter modified Kneser-Ney [16] smoothing (KN). For better performance, we apply n-gram weighting with document entropy features [4] to the component models to de-emphasize out-of-domain n-grams (KN+W). We interpolate these component models built from Textbook and Lectures using linear interpolation (LI) and count merging (CM) [7]. Overall, the KN and KN+W model configurations have 1 and 7 parameters, respectively.

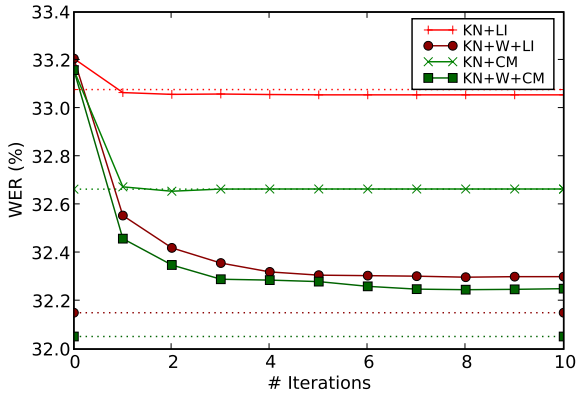

Fig. 1. Average test set WER vs. adaptation iteration. The 0^{th} iteration corresponds to the unadapted model. The dotted lines correspond to the oracle WERs obtained by tuning the model on the reference transcript.

2.2. Recognition Hypotheses

In scenarios where no matched training data is available, we can perform multiple recognition passes and use the 1-best recognition hypotheses from the previous pass as the development set for tuning the LM parameters. In Figure 1, we plot the WERs for various LM configurations over 10 such iterations. The 0^{th} iteration corresponds to estimating the LM using the default parameter values. As baselines, we also include, as dotted lines, the oracle WERs obtained by tuning the model parameters directly on the reference transcript.

As observed in previous works, count merging outperforms linear interpolation. Without n-gram weighting, the WER for both interpolation techniques converges in a single iteration. The errors in the recognition hypotheses appear to have negligible effect on the tuned performance of these 1-parameter models. [1]

Overall, applying n-gram weighting to the individual LM components significantly improves the performance of the resulting models by introducing additional tuning parameters. Unlike the 1-parameter models, the WER of the n-gram weighted models does not converge until about the 4^{th} iteration, with the first iteration achieving only about 75% of the total reduction. Furthermore, it is much more sensitive to errors in the development set, with a gap of 0.2% between the best unsupervised and the oracle WERs at about 30% development set WER.

[1] The difference between the best unsupervised and the oracle WERs for KN+LI is not statistically significant and can be attributed to optimizing for minimum development set perplexity instead of WER.

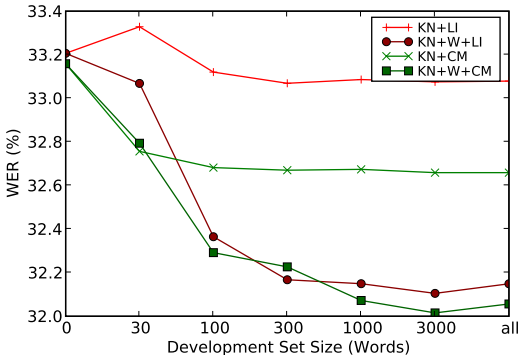

Fig. 2. Average test set WER vs. development set size for various LM configurations. 0 corresponds to the unadapted model. all corresponds to approximately 10,000 words.

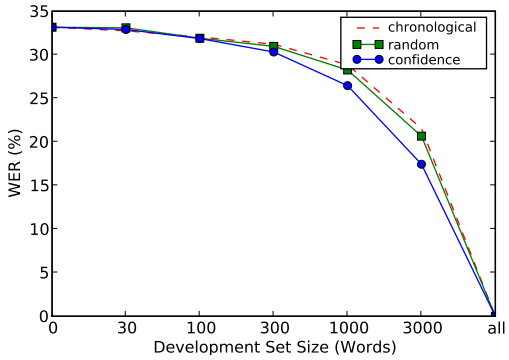

Fig. 3. Average effective test set WER vs. development set size for the KN+W+CM configuration using various utterance selection strategies.

In some configurations, unsupervised adaptation has previously been observed to degrade performance after a few iterations due to the reinforcement of previous recognition errors [7]. By using the recognition hypotheses only to tune the model parameters, the above models appear to be immune to such overfitting.

2.3. User-Transcription

Although human transcription of the entire lecture is often impractical, content providers and end-users are often motivated to help improve the recognition accuracy. Instead of using inaccurate recognition hypotheses, we may be able to obtain transcripts of select utterances from the users for the development set. Unlike general transcriptionists, users of the speech application are also more likely to correctly transcribe the technical jargon found in the target lectures. In Figure 2, we plot the performance of various LM configurations trained with increasing amounts of development set data. Specifically, we incrementally add the reference transcripts of random utterances from the target lecture until we have obtained the desired minimum number of words.

As expected, increasing the development set size improves the LM performance. For simple linear interpolation and count merging, the WER converges after only 100 words. Although applying n-gram weighting reduces the WER by up to 1.0% absolute, it takes about 10 times more development set data before we observe convergence.

Compared with iterative unsupervised adaptation using the recognition hypotheses, we are able to achieve better recognition performance with only 300 words of transcribed development set data, or 3% of the target lecture, for all model configurations. With appropriate transcription tools that support re-dictation of the target utterance, easy correction for speech recognition errors [17], and a streamlined text entry interface [18], we expect most users to be able to transcribe 300 words within 15 minutes.

2.4. Utterance Selection

Since the application often have the opportunity to preprocess the target lecture prior to presenting the utterances to the user for transcription, we can consider various utterance selection strategies to best utilize the user transcriptions. The most natural approach from the user perspective is to transcribe the utterances in chronological order, as they are easier to comprehend in context. However, given the topic shifts that frequently occur within a lecture [19], random selection may result in a development set that better represents the target lecture.

As user transcribed utterances from the target lecture do not need to be processed by the speech recognizer, which adds to the overall WER, another strategy is to select utterances that are most likely to yield recognition errors. In our implementation, we compute the confidence of an utterance as the average of the 1-best word posterior probabilities in a normalized word lattice [10] and present the utterances in order of increasing confidence.

Since transcribed utterances do not contribute to the WER, we need to remove errors corresponding to transcribed utterances when computing the effective WER. Thus, when all utterances are transcribed, the effective WER is 0. In Figure 3, we plot the effective WER against the minimum number of words transcribed for the chronological, random (averaged over 3 trials), and lowest confidence utterance selection strategies. As predicted, random selection generally yields lower WER than chronological selection. Although transcribing utterances in order of lowest confidence may result in a development set less representative of the overall lecture, in this task, the benefit of not accumulating errors from these low confidence utterances outweigh the effects of the mismatch. If users can transcribe low confidence utterances with the same effort as high confidence ones, utterances to be transcribed should be selected in order of increasing confidence to minimize the effective WER.

3. CONCLUSION & FUTURE WORK

In this work, we demonstrated the effectiveness of various techniques to tune multi-parameter LMs when matched training data is unavailable. When interpolating Textbook and Lectures using count merging with n-gram weighting, optimizing the parameters using the recognition hypotheses as a development set achieves around 80% of the 1.1% oracle WER reductions obtained with the reference transcript. Whereas single-parameter models generally converge in a single adaptation iteration, more sophisticated models often require iterative adaptation to achieve the best performance.

In supervised settings with the same LM configuration, 300 words of user transcription is sufficient to outperform unsupervised adaptation. By selecting the utterances with the lowest confidence for transcription, we can further reduce the effective transcription WER by another 0.6%.

For future work, we plan to present the efficient lattice rescoring data structure and algorithms that enable us to practically conduct the above experiments. We also would like to explore using both the user transcriptions and recognition hypotheses for LM adaptation. In addition to using the hypotheses for text selection, we hope to examine the use of word-level confidence scores as n-gram weighting features to train a component model from the hypotheses.

In real-world applications, users are often motivated to help improve the recognition accuracy. In addition to providing utterance transcriptions, they can also contribute by gathering additional relevant textual material, crucial to reducing out-of-vocabulary words, for LM training. Traditional speech and natural language processing research often focus on particular problems in isolation, without sufficient regard to the application contexts in which these problems occur. In addition to investigating models and algorithms that improve system performance given fixed data resources, we need to explore new ways in which users interact with and contribute to the system in order to build more effective natural user interfaces for next-generation applications.

Acknowledgments

We would like to thank Jay Patel for assistance with preliminary experiments and the reviewers for their constructive feedback. This research is supported in part by the T-Party Project, a joint research program between MIT and Quanta Computer Inc.

4. REFERENCES

[1] James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang, "Analysis and processing of lecture audio data: Preliminary investigations," in *Proc. HLT-NAACL Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, Boston, MA, USA, 2004.

[2] Jerome R. Bellegarda, "Statistical language model adaptation: Review and perspectives," *Speech Communication*, vol. 42, no. 1, pp. 93–108, 2004.

[3] Bo-June (Paul) Hsu, "Generalized linear interpolation of language models," in *Proc. ASRU*, Kyoto, Japan, 2007.

[4] Bo-June (Paul) Hsu and James Glass, "N-gram weighting: Reducing training data mismatch in cross-domain language model estimation," in *Proc. EMNLP*, Honolulu, Hawaii, USA, 2008.

[5] Hiroaki Nanjo and Tatsuya Kawahara, "Unsupervised language model adaptation for lecture speech recognition," in *Proc. SSPR*, Tokyo, Japan, 2003.

[6] Gokhan Tur and Andreas Stolcke, "Unsupervised language model adaptation for meeting recognition," in *Proc. ICASSP*, Honolulu, Hawaii, USA, 2007.

[7] Michiel Bacchiani and Brian Roark, "Unsupervised language model adaptation," in *Proc. ICASSP*, Hong Kong, China, 2003.

[8] Thomas Niesler and Daniel Willett, "Unsupervised language model adaptation for lecture speech transcription," in *Proc. Interspeech*, Denver, Colorado, USA, 2002.

[9] Langzhou Chen, Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. ICASSP*, Hong Kong, China, 2003.

[10] Giuseppe Riccardi and Dilek Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[11] James Glass, Timothy J. Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[12] James Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech & Language*, vol. 17, no. 2-3, pp. 137–152, 2003.

[13] Hung-An Chang and James Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," 2008, submitted to *ICASSP*.

[14] Timothy J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. Interspeech*, Pittsburgh, PA, USA, 2006.

[15] Bo-June (Paul) Hsu and James Glass, "Iterative language model estimation: Efficient data structure & algorithms," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[16] Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," in *Technical Report TR-10-98*. Computer Science Group, Harvard University, 1998.

[17] David Huggins-Daines and Alexander I. Rudnicky, "Interactive ASR error correction for touchscreen devices," in *Proc. ACL*, Columbus, Ohio, USA, 2008.

[18] Rony Kubat, Philip DeCamp, Brandon Roy, and Deb Roy, "TotalRecall: Visualization and semi-automatic annotation of very large audio-visual corpora," in *Proc. ICMI*, Antwerp, Belgium, 2007.

[19] Bo-June (Paul) Hsu and James Glass, "Style & topic language model adaptation using HMM-LDA," in *Proc. EMNLP*, Sydney, Australia, 2006.