Evaluating Unsupervised Language Model Adaptation Methods for Speaking Assessment

Shasha Xie
Microsoft
1020 Enterprise Way
Sunnyvale, CA 94089
shxie@microsoft.com

Lei Chen
Educational Testing Service
600 Rosedale Rd
Princeton, NJ
LChen@ets.org

Abstract

In automated speech assessment, adaptation of language models (LMs) to test questions is important to achieve high recognition accuracy However, for large-scale language tests, the ordinary supervised training, which uses an expensive and time-consuming manual transcription process, is hard to utilize for LM adaptation. In this paper, several LM adaptation methods that require either no manual transcription process or just a small amount of transcriptions have been evaluated. Our experiments suggest that these LM adaptation methods can allow us to obtain considerable recognition accuracy gain with no or low human transcription cost.

Index Terms: language model adaptation, unsupervised training, Web as a corpus

1 Introduction

Automated speech assessment, a fast-growing area in the speech research field (Eskenazi, 2009), typically uses an automatic speech recognition (ASR) system to recognize spontaneous speech responses and use the recognition outputs to generate the features for scoring. Since the recognition accuracy directly influences the quality of the speech features, especially for the features related to word entities, such as those measuring grammar accuracy and vocabulary richness, it is important to use ASR systems with high recognition accuracy.

Adaptation of language models (LMs) to test responses is an effective method to improve recognition accuracy. However, it is difficult to only use the ordinary supervised training to adapt LMs to test questions. First, for high-stake tests administered globally, a very large pool of test questions have to be used to strengthen the tests' security and validity. Since a large number of test questions have many possible answers for each question, a large set of audio files needs to be transcribed to cover response content. Second, due to time and cost constraints, it may not be practical to have a pre-test to collect enough speech responses for adaptation purposes. Therefore, it is important to pursue other methods to obtain LM adaptation data in a faster and lower-cost way than the ordinary supervised training.

As we will review in Section 2, some promising technologies, such as *unsupervised training*, *active learning*, and *LM adaptation based on Web data*, have been utilized in broadcast news recognition, dialog system, and so on. In this paper on the LM adaptation task used in automated speech scoring systems, we will report our experiments to obtain LM adaptation data in a faster and more economical way that requires little human involvement. To our knowledge, this is the first such work reported in the automated speech assessment area.

The rest of the paper is organized as follows: Section 2 reviews the related previous research results; Section 3 describes the English test, the data used in our experiments, and the ASR system used; Section 4 reports the experiments of different methods we tried to obtain LM adaptation data; Section 5 discusses our findings and plans for future research.

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 288–292,
Atlanta, Georgia, June 13 2013.

2 Previous Work

Unsupervised training is the method of using untranscribed audio to adapt a language model (LM). An initial ASR model (seed model) is used to recognize the untranscribed audio, and the obtained ASR outputs are used in the follow-up LM adaptation. (Chen et al., 2003) utilized unsupervised LM adaptation on broadcast news (BN) recognition. The unsupervised adaptation method reduces the word error rate (WER) by 2% relative to using the baseline LM. (Bacchiani and Roark, 2003) reported that unsupervised LM adaptation provided an absolute error rate reduction of 3.9% over the un-adapted baseline performance by using 17 hours of untranscribed adaptation data. This was 51% of the 7.7% adaptation error rate reduction obtained by using an ordinary supervised adaptation method.

Active learning is used to reduce the number of training examples to be annotated by automatically processing the unlabeled examples and then selecting the most informative ones with respect to a given cost function. (Riccardi and Hakkani-Tur, 2003; Tur et al., 2005) proposed using a combination of unsupervised and active learning for ASR training to minimize the workload of human transcription. Their experiments showed that the amount of labeled data needed for a given recognition accuracy can be reduced by 75% when combining these two training approaches.

A recent trend in Natural Language Processing (NLP) and speech recognition research is utilizing Web data to improve the LMs, especially when in-domain training material is limited. (Ng et al., 2005) investigated LM topic adaptation using Web data. Experiments in recognizing Mandarin telephone conversations showed that use of filtered Web data leads to a 7% reduction in the character recognition error rate. (Sarikaya et al., 2005) used Web data to adapt LMs used in a spoken dialog system. From a limited in-domain data set, they generated a series of search queries and retrieved Web pages from Google using these queries. In their recognition experiment done on a dialog system, they achieved a 5.2% word error reduction by using the Web data, compared to a baseline LM trained on 1700 in-domain utterances.

3 Test, Data, and ASR

Our in-domain data was from The Test of English for International Communication, TOEIC®, which tests non-native English speakers' basic speaking ability required in international business communications. In our experiments, we focused on *opinion* testing questions. An example question is: "*Do you agree with the statement that a company should only hire experienced employees? Use specific reasons to support your answer*".

A state-of-the-art HMM LVCSR system, which was provided by a leading ASR vendor, was used in our experiments. It contains a cross-word tri-phone acoustic model (AM) and a combination of bi-gram, tri-gram, and up to four-gram LMs. The AM and LM are trained by supervised training from about 800 hours of audio and manual transcriptions of non-native English speaking data collected from the Test Of English as a Foreign Language (TOEFL®). TOEFL® is targeted to assess test-takers' ability to use English to study in an institution using English as its primary teaching language. Speaking content from TOEFL® data is quite different from the content shown in TOEIC® data. When testing this recognizer on a held-out evaluation set extracted from the TOEFL® test, a word error rate (WER) of 33.0% [1] is observed. This recognizer was used as the *seed* recognizer in our experiments.

4 Experiments

We collected a set of audio responses from the TOEIC® test, focusing on opinion questions. This data set was randomly selected from different first-language (L1) and English speaking proficiency levels. Then, these audio files were manually transcribed. In our experiments, 1470 responses were used for LM adaptation and the remaining 184 responses were used to evaluate speech recognition

[1] ASR on non-native speech is more difficult than on native speech for various reasons (Livescu and Glass, 2000). However, a high WER does not rule out the possibility of using ASR outputs for automated scoring, especially when relying on delivery related features. For example, (Chen et al., 2009) shows that several pronunciation features' contributions for assessment, measured as Pearson correlations between the featuers and human scores, only drop about 10% to 20% when using ASR outputs with a WER as high as 50% compared to using human transcriptions.

accuracy. When using the seed recognizer without any adaptation, the WER on the evaluation set is 42.8%, which is much higher than the accuracy achieved on the TOEFL® data (33.0%). Using the ordinary supervised training, adapting LMs using these 1470 manual transcriptions, the WER is reduced to 34.7%, close to the performance on the in-domain TOEFL® data. Note that a fixed dictionary with a vocabulary size of about 20,000 words, which in general is much larger than the vocabulary mastered by non-native test takers, was used in our experiment.

4.1 Unsupervised LM adaptation

Using the seed recognizer trained on the TOEFL® data, we recognized 1470 adaptation responses and selected varying amounts of ASR outputs for LM adaptation. From ASR outputs of all responses, we selected the responses with high confidence scores estimated by the seed recognizer so that we could use the ASR outputs with higher recognition accuracy on the LM adaptation task. We used two methods to measure the confidence score for each response from word-level confidence scores. First, we took the average of all word confidence scores a response contains, as shown in Equation 1.

$$Conf_{perWord} = \frac{1}{N}\sum_{i=1}^{N} conf(w_i) \qquad (1)$$

where $conf(w_i)$ is the confidence score of word, w_i. The other method we used considers each word's duration, as shown in Equation 2.

$$Conf_{perSec} = \frac{\sum_{i=1}^{N} d(w_i) * conf(w_i)}{\sum_{i=1}^{N} d(w_i)} \qquad (2)$$

where $d(w_i)$ is the duration of w_i.

In Figure 1, we showed the WER after running unsupervised LM adaptation, where the adaptation responses were selected if they had high word-based ($Conf_{perWord}$) or duration-based ($Conf_{perSec}$) confidence scores. The data sizes used for adaptation vary from 0% (without any adaptation) to 100% (using all adaptation data). We observe continuous reduction of WER when using more and more adaptation data. Selecting responses by the word-based confidence scores performs a little better than the selection method based on the confidence scores normalized by corresponding word durations. However, there is no significant difference between these two selection criteria.

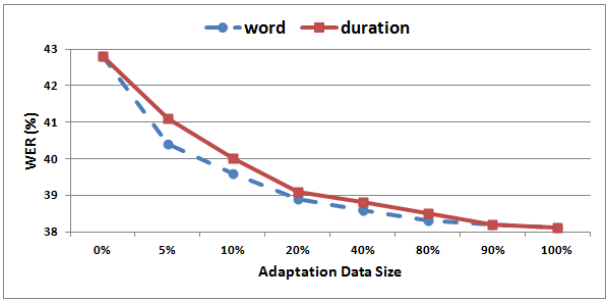

Figure 1: Unsupervised LM adaptation performance using different sizes of development set data.

ASR accuracy may vary within each response. Therefore, instead of using entire responses, we also explored using smaller units for LM adaptation. All of the ASR outputs were split into word sequences with fixed lengths (10-15 words), and the ones with higher per-word confidence scores ($Conf_{perWord}$) were extracted for model adaptation. Our experiment shows that using word-sequence pieces rather than entire responses leads to a faster WER reduction. When only using 5% of the adaptation data, we obtained 3.5% absolute WER reduction compared to the baseline result without adaptation. Note that we only obtained 2.5% absolute WER reduction when using entire responses in adaptation.

4.2 Web data LM adaptation

Given around 40% WER when using our seed ASR, unsupervised learning faces the issue that many recognition errors were included in model adaptation. Can we find another source to obtain LM adaptation inputs with fewer errors? To address this question, we explored building a training corpus from Web data based on test questions. We used BootCat (Baroni and Bernardini, 2004), a corpus building tool designed to collect data from the Web, to collect our LM adaptation data. Based on test prompts in the TOEIC® test, we manually generated search queries. After receiving the search queries, the BootCat tool searched the Web using the Microsoft Bing search engine. Then, top-ranked

Web pages were downloaded and texts on these Web pages were extracted. We examined the Web search results (including URLs and texts) returned by the BootCat tool. The returned Web data has varied matching rates among these prompts and are generally noisy.

By using only the default setup provided by the BootCat tool, we collected 5312 sentences in total. After a simple text normalization, we used the obtained Web data for LM adaptation, and the WER on the evaluation data was 38.5%. This WER result is a little higher than the WER result achieved by unsupervised LM adaptation (38.1%). Without transcribing any response from test-takers, the language model adaptation using Web data already helps to improve recognition accuracy. Then, we tried using both the Web data and the ASR hypotheses for adaptation, and we can further decreased the WER to 37.6%. This is lower than using the two LM adaptation data sets separately.

4.3 Semi-supervised approaches for LM adaptation

For semi-supervised LM adaptation, we replaced the speech responses of lower confidence scores with their corresponding human transcripts. We hoped that by using the responses with high confidence scores together with a small amount of human transcripts, we could get better performance by introducing less noise during adaptation. We set different thresholds for selecting the low confidence responses and replacing them with human transcripts. We find that just manually transcribing a limited amount of audio data gives us further WER reduction, compared to using unsupervised learning. After transcribing just 100 responses, 6.8% of 1470 responses in the adaptation data set, semi-supervised learning can achieve 61.73% of the WER reduction (8.1%) obtained by using the ordinary supervised training that requires transcription of all 1470 responses.

4.4 Discussion

In Table 1, we compared the performance of all the adaptation methods mentioned in this paper, including two unsupervised methods adapted using the ASR hypotheses and "related" Web data, and one semi-supervised method [2], replacing the ASR hypotheses of lower confidence scores with their corresponding human transcripts. For a convenient comparison, we also include the baseline (without LM adaptation) and the result of using the supervised adaptation. All the proposed unsupervised/semi-supervised methods can significantly improve the ASR performance compared to the baseline result. For projects with time limits, we can use these unsupervised/semi-supervised methods to help us get relatively good ASR outputs.

Table 1: The WER on the evaluation set using different LM adaptation methods.

baseline	*unsupervised*			*semi*	*super.*
	ASR	*Web*	*ASR&Web*		
42.8	38.1	38.5	37.6	37.8	34.7

5 Conclusions and Future Work

In this paper, we reported our experiments in applying several LM adaptation methods to automated speech scoring systems that require few, if any, human transcripts, which are expensive and slow to obtain for large-sized adaptation data sets. The unsupervised training (using ASR transcriptions from a seed ASR system) clearly shows higher accuracy than a ASR system without any domain adaptation. We also used test questions to collect related texts from Web. Even though such Web data may be noisy and its relatedness to real test responses is not always guaranteed, text data collected from the Web is helpful to adapt LMs to better fit the responses to test questions. To better cope with recognition errors brought on by using the unsupervised training method, we proposed using human transcriptions on a small amount of poorly recognized responses. Using such little human involvement further helps to obtain a lower WER. Therefore, based on the experiments described in this paper, we conclude that these novel LM adaptation methods provide promising solutions to let us skip the ordinary supervised training for LM adaptation tasks frequently used in automated speech scoring.

[2]The semi-supervised result was from replacing 100 low-confidence responses with human transcripts.

The reported experiments in this paper were conducted on a limited-size data set. We plan to increase the testing data to a larger size and hope to cover more types of test questions and spoken tests. In addition, we plan to investigate how to automatically generate Web search queries based on test questions.

References

M. Bacchiani and B. Roark. 2003. Unsupervised language model adaptation. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*.

M. Baroni and S. Bernardini. 2004. BootCaT: bootstrapping corpora and terms from the web. In *Proceedings of LREC*, volume 2004, page 13131316.

L. Chen, J. L Gauvain, L. Lamel, and G. Adda. 2003. Unsupervised language model adaptation for broadcast news. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*.

L. Chen, K. Zechner, and X Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *NAACL-HLT*.

M. Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

K. Livescu and J. Glass. 2000. Lexical modeling of non-native speech for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1683–1686.

T. Ng, M. Ostendorf, M. Y Hwang, M. Siu, I. Bulyko, and X. Lei. 2005. Web-data augmented language models for mandarin conversational speech recognition. In *Proc. ICASSP*, volume 1.

G. Riccardi and D. Z Hakkani-Tur. 2003. Active and unsupervised learning for automatic speech recognition. In *Proc. 8th European Conference on Speech Communication and Technology*.

R. Sarikaya, A. Gravano, and Y. Gao. 2005. Rapid language model development using external resources for new spoken dialog domains. In *Proc. ICASSP*, volume 1, pages 573–576.

G. Tur, D. Hakkani-Tur, and R. E Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186.