An Error-Corrective Language-Model Adaptation For Automatic Speech Recognition

Minwoo Jeong, Jihyun Eun, Sangkeun Jung, Gary Geunbae Lee

Department of Computer Science and Engineering,
Pohang University of Science & Technology (POSTECH)
{stardust, tigger, hugman, gblee}@postech.ac.kr

10.21437/Interspeech.2005-12

Abstract

We present a new language model adaptation framework integrated with error handling method to improve accuracy of speech recognition and performance of spoken language applications. The proposed error corrective language model adaptation approach exploits domain-specific language variations and recognition environment characteristics to provide robustness and adaptability for a spoken language system. We demonstrate some experiments of spoken dialogue tasks and empirical results which show an improvement of the accuracy for both speech recognition and spoken language understanding.

1. Introduction

A spoken language interface is often required in many application environments, such as mobile information retrieval, car navigation systems, and ubiquitous computing, but the low speech recognition performance makes it difficult to extend its application to new fields. In a spoken language system to provide practical interaction between human and machine, one of the major problem is how to recover decreasing application level performance due to incomplete outputs in speech recognition. Especially, accuracy of automatic speech recognition (ASR) is severely affected according to the recognition task changes. In a new recognition task, the lexical, syntactic, or semantic characteristics of the utterance have different distributions in comparison with the training corpus. To recover this mismatch and improve abilities of spoken language systems, speech recognizer should be adapted to a specific recognition domain.

Various language model adaptation approaches were investigated [1]. The most widespread approaches of adaptive language modeling use the N-best re-scoring technique, which is an efficient ad-hoc method to adapt to new observed data in the recognition task. Typically, N-best re-scoring assumes that the correct answer sentence should be always included in the N-best hypotheses. But, if a speech recognizer is not enough to provide correct words or lattices due to extremely noisy environments or abnormal characteristics of speakers, even the N-best hypotheses cannot cover the correct answer sentences in many of the cases. To overcome this problem, we propose a new language model adaptation method to incorporate channel characteristics of the domain environment as well as various higher-level linguistic knowledge.

2. Error Corrective Language Model Adaptation

For robust spoken language systems, we present a general unified framework to correct first-pass ASR errors and to adapt the linguistic variations of the recognition task. The adaptation framework can handle ASR errors and also combine high-level linguistic knowledge in a uniform manner.

2.1. Adaptation Framework

We can consider three different data: a large background corpus B, collected from related or somewhat different tasks, a small channel adaptation data A', output of ASR in the current task, and a small linguistic feature corpus A'', relevant to the current recognition task (quite similar or equal to A'). Then, the entire adaptation data $A = A' \bigcup A''$.

We assume that domain independent ASR would produce erroneous output sequence (A'), and adaptation procedure would find the new word sequence with adapted language model distribution (A''). So, the problem of error corrective language model (ECLM) adaptation can be stated in this model as follows. For an input sentence, $o = o_1, o_2, \ldots, o_n$, which is produced as the output sequence of ASR, find the best word sequence, $w^* = w_1, w_2, \ldots, w_n$, that maximizes the posterior probability $P_A(w|o)$. Then, applying Bayes' rule and dropping the constant denominator, we can rewrite it as:

$$w^* = \arg\max_w P_A(o|w) \cdot P(w). \quad (1)$$

At this point, we have a noisy channel (or source-channel) model for ECLM adaptation, with two components, the channel model $P_A(o|w)$ estimated by A' and the language model $P(w)$ initially generated by corpus B and later adapted by A''. This model can be viewed as a maximum a posteriori (MAP) adaptation strategy with different parameterizations of the prior distribution or as an optimized model approach to minimize error rate in post-processing on recognizer output [6].

2.2. Channel Modeling

The conditional probability, $P_A(o|w)$ reflects the channel characteristics of the ASR environment. If we assume that the output word sequences produced by ASR are independent of one another, we have the following formula:

$$P_A(o|w) = P_{A'}(o_1, \ldots, o_n|w_1, \ldots, w_n) = \prod_{i=1}^{n} P_{A'}(o_i|w_i) \quad (2)$$

where channel model is based on data A', so $P_A(o|w) = P_{A'}(o|w)$. However, this simple one-to-one model is not suit-

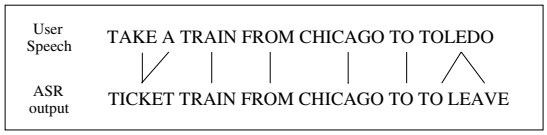

Figure 1: An example of a channel modeling

able for handling split or merged errors, which frequently appear in an ASR output, because errors influence the surrounding words, and there is a context dependency in the error word sequence. For example, Fig. 1 shows a split or a merged error problem (example from [6]). This problem was firstly addressed in a statistical machine translation (MT) community [2]. Following [2], we refer to the k-number of post-channel words o_i produced by a pre-channel word w_i as a fertility with probability $P_{A'}(f : k|w_i)$. We can simplify the fertility model of IBM statistical MT model-4, and allow the fertility within 2 windows such as $P_{A'}(o_{i-1}, o_i|w_i) \cdot P_{A'}(f : 2|w_i)$ for two-to-one channel probability, and $P_{A'}(o_i|w_i, w_{i+1}) \cdot P_{A'}(f : 1/2|w_i) \cdot P_{A'}(f : 1/2|w_{i+1})$ for one-to-two channel probability. So, the fertility model can deal with (TICKET, TAKE A) or (TO LEAVE, TOLEDO) substitution in the example of Fig. 1.

To train the channel model, we need a training data consisting of $\{w, o\}$ pair data A' which are manually transcribed strings and ASR outputs. Also, we align the pair based on minimizing the edit distance between w_i and o_i by dynamic programming. We can then calculate the probability of each substitution $P_{A'}(o_i|w_i)$ by maximum-likelihood estimation (MLE). If adaptation data A' is not enough to estimate channel model parameters, we will have $P_{A'}(o_i|w_i) = 0$, and less or no candidates to generate the corrected sentences. To overcome this unseen event, we smooth the channel model with modified Kneser-Ney discounting method [3].

2.3. Language Modeling

The probability $P(w)$ is given by the language model and plays a pivotal role of the prior. The distribution $P(w)$ can be defined using *n*-gram, structured language model, or any other tool in statistical language modeling. For adaptive language modeling, $P(w)$ should be combined with the background corpus B and the adaptation corpus A''. Linear interpolation is the simplest way to merge the models. Given estimates for word w_i denoted by $P_{A''}(w_i|h_i)$ and $P_B(w_i|h_i)$, a merged model $P(w_i|h_i)$ can be expanded as follows:

$$P(w_i|h_i) = (1 - \lambda)P_{A''}(w_i|h_i) + \lambda P_B(w_i|h_i) \qquad (3)$$

where $0 \leq \lambda \leq 1$ is the interpolation coefficient, and $P(w) = \prod_i P(w_i|h_i)$. In language model adaptation, the corpus B is used in ASR language model to generate hypotheses via first pass recognition, and the adaptation data A'' will be collected by domain-specific text or utterances to correct and re-score the hypotheses.

We used two kinds of language models: a word *n*-gram model and the whole-sentence maximum entropy language model (WSME-LM) [7]. At the first level, a word *n*-gram model was used for capturing local dependencies and for rapid processing. The WSME-LM was used at the second level to capture long-distance dependency and higher-level linguistic phenomena, and to re-score the N-best hypotheses produced by the first level adaptation.

3. Re-ranking using Linguistic Knowledge

3.1. Re-ranking Model

The *n*-gram models lack the knowledge to overcome the low accuracy of the current spoken language systems. To exploit multiple knowledge sources (e.g. part-of-speech tag, syntactic dependency, or semantic information), the adaptation data A'' is used to extract specific linguistic information about different aspects of the mismatch between training and recognition conditions. Because the WSME-LM can combine *n*-gram features and other higher level linguistic knowledge, we used WSME-LM to re-rank the error-corrected (through channel model) baseline hypotheses.

Using the WSME-LM is straightforward and computationally trivial, because the universal normalization constant is not dependent on history and needs not be calculated [7]. However, it is impossible to compute the expectations and normalization factor directly, which requires a summation over all possible sentences. Instead, they can be approximated using a set of representative samples generated by sampling from the distribution.

3.2. Sampling Method

Several Markov chain monte carlo (MCMC) sampling techniques have been investigated for WSME-LM training [7]. Importance sampling was used in our experiments. In importance sampling, we are able to sample from some other distribution that approximates $P(s)$, whose probability densities are proportional to base probability $P_b(s)$. We base our estimates on a sample $\{s_1, \ldots, s_N\}$, generated from the distribution $P_b(s)$. We can then estimate the expectation of $E_p f_i$ with respect to the distribution $P(s)$ by

$$E_p f_i \approx \frac{\sum_{j=1}^{N} a_j \cdot f_i(s_j)}{\sum_{j=1}^{N} a_j}. \qquad (4)$$

where an importance weight is $a_j = \frac{P(s_j)}{P_b(s_j)}$ for each s_j. The accuracy of approximated $E_p f_i$ depends on the variability of the importance weights. For this importance sampling to work well, the distribution $P_b(s)$ must be a fairly good approximation to the one defined by $P(s)$. The initial trigram distribution $P_0(s)$ can be a good choice for efficient sampling.

Generating sentences from a trigram model can be done quite efficiently. For efficient processing, vocabulary is restricted by domain vocabulary (including common words), and maximum sentence length is restricted to constant value (e.g. 10 to 40). End-of-sentence marker </s> can stop expanding the sentence. So, a set of generated sentences, $\{s_1, \ldots, s_N\}$, is a highly probable sentence set in the sampling domain.

4. Experiments

4.1. Data Sets

We have evaluated our ECLM adaptation method using the two standard dialogue corpora for English and Korean: CU-Communicator travel data [1] (CU-Data) and Korean tele-banking dialogue data [2] (TeleBank-Data). The CU-Data was collected

[1] CU-Data made by University of Colorado is an over-the-telephone spoken dialogue corpus to develop a dialogue system for accessing live airline, hotel and car rental information.

[2] TeleBank-Data is a Korean spoken dialogue corpus which is provided by Sogang University.

in 461 days and consists of 2211 dialogues or 38,408 utterances in total. After cleaning up the data [3], the test set consists of 12,961 utterances. The test set of Telebank-DATA consists of 4,558 utterances. Because TeleBank-Data set was collected by restricted scenario and vocabulary, the data is small and the speech recognition accuracy of this data is higher than that of the CU-Data.

For CU-Data, we used open source speech recognizer Sphinx2 and open source CMU-Communicator Feb-2000 acoustic model [4] which was trained with semi-continuous density model. We use TDT3 for text corpus which is a collection of News articles and contains 177K sentences for background language model. For TeleBank-Data, we made an HTK-based Korean speech recognizer which was trained by mel-frequency cepstrum coefficients (MFCC) 39 dimensional feature vectors. We use MATEC [5] text corpus which consists of 34K sentences from novel and news articles for background language model. For each test sentence, we produced 100-best lists with these speech recognizers.

And we divided each test set into 10 different sets, and evaluated the results of 10-fold cross validation for all our experiments.

4.2. Model Training and Feature Selection

We modeled 10-different channel models for cross validation evaluation with each test data. We got each 44,546 and 19,796 channel model parameters of CU-Data and TeleBank-Data on average. Using maximum of 2 fertilities, we got each 964 and 407 fertility model parameters of each data on average.

Then, we constructed two background trigram models on background corpora: TDT3 and MATEC articles. The background trigram models were applied with modified Kneser-Ney smoothing [3], and were interpolated with adaptation data A''. For adaptation data A'', we used 10K utterances for CU-Data and 4K utterances for TeleBank-Data.

Using the trigram models, we generated 100K sample sentences for each test set, and used them to train WSME-LM with importance sampling method. We trained WSME-LM with GIS algorithm and Gaussian prior, and ended up with 55,886 features of CU-Data and 45,677 features of TeleBank-Data with 5-frequency cutoff. We used several features for our ECLM adaptation as in the following:

- Distance-k n-gram
  Distance-k n-gram (d-k-n-gram) is an n-gram with k words back to the word to be predicted. For example, d-2-bigram predicts w_i based on w_{i-2} and d-2-trigram predicts w_i based on w_{i-3}, w_{i-2}.
- Syntactic features
  Using part-of-speech tagging and parsing, we automatically added syntactic features to combine syntax with lexical features. For instance, features can be a pair of word w with part-of-speech tag t, chunk c, or head word h such as $\{w_i, t_{i-1}\}$, $\{w_i, c_{i+1}\}$, $\{w_i, h_i\}$.
- Sentence length and others
  We added some supra-structure of sentences such as sentence length, person, and dialogue features.

[3] We removed most of the single word responses such as "Yes, please" or "No" which were not very useful to evaluate the spoken language understanding (SLU) performance.

[4] Available at http://www.speech.cs.cmu.edu/Communicator

[5] MATEC (Morphological Analysis and Tagging Evaluation Conference): An ETRI-supported morphological tagging contest in Korean.

Table 1: Results of ECLM adaptation on CU-Data and TeleBank-Data. WER with reduction rate on baseline recognition and ECLM adaptation with two different language models.

	CU-Data	TeleBank-Data
baseline (trigram)	31.34%	18.71%
ECLM (trigram)	29.75% (-5.07%)	14.10% (-24.64%)
ECLM (+WSME-LM)	28.81% (-8.07%)	12.21% (-34.74%)

4.3. Empirical Results of Automatic Speech Recognition

Table 1 presents the experiment results of ECLM adaptation in CU-Data and TeleBank-Data evaluation. The word error rate (WER) of the baseline[6] ASR system is 31.34% on the utterances in CU-Data. Using ECLM adaptation with only the initial trigram model and with WSME-LM, we achieved a 5.07% and 8.07% error reduction rate for WER. And, the WER of the baseline is 18.71% in TeleBank-Data. We achieved a 24.64% and 34.74% error reduction rate for WER via ECLM adaptation.

CU-Data can be divided into 25 sections which are assembled by date. One section of CU-Data consists of utterances collected by a period of one month. Using NIST Speech Recognition Scoring Toolkit[7], we plot the monthly WER of CU-Data sorted with WER in Fig 2. Unlike CU-Data, TeleBank-Data is not classified by the date but by the speaker. We draw the WER of each 113 speakers again sorted with WER in Fig 3. As shown in Fig. 2 and 3, ECLM adaptation is very significant to improve the accuracy of the baseline speech recognition system for almost all months and all speakers.

We performed the standard NIST Matched Pairs Sentence Segment Word Error Test (MAPSSWE) as a statistical significance test [5] for each test set and found that:

- the WER improvement of the ECLM adaptation with trigram over the baseline model is significant ($p < 0.001$),
- that of the ECLM adaptation with WSME-LM over the baseline model is also significant ($p < 0.001$), and
- that of the ECLM adaptation with WSME-LM over the ECLM adaptation with trigram is also significant ($p < 0.001$).

4.4. Empirical Results of Spoken Language Understanding

To validate our approach in spoken language understanding (SLU) applications, we conducted an experiment on a SLU system. The goal of our SLU system is to extract the meaning from the recognized user's utterances based on the semantic frame. A reference semantic frame (or template) is a well-formed structure of the extracted information consisting of slot/value pairs. To extract semantic frames from user's utterance inputs, we used a linear-chain conditional random fields (CRF) [4]. The basic n-gram, d-k-n-gram, and part-of-speech tags were used as observed features for our CRF-based SLU system. We defined 22 semantic class slots to be extracted from CU-Data and 17 slots from TeleBank-Data.

[6] The result of baseline is 1-best output of ASR, whose trigram is interpolated with background and domain-specific corpus (as in Eq. 3).

[7] Available at http://www.nist.gov/speech/tools/

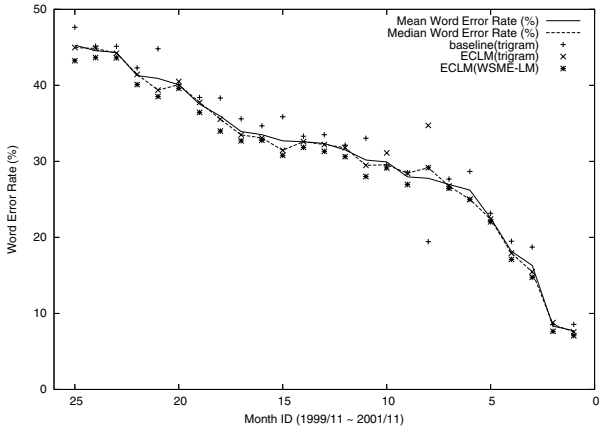

Figure 2: Monthly error rate in CU-Data

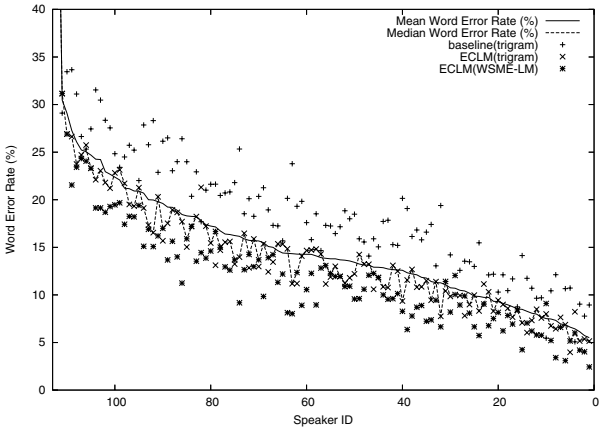

Figure 3: Each speaker's error rate in TeleBank-Data

Table 2: SLU evaluation results of CU-Data

	Acc	Prec	Rec	F1
baseline (trigram)	53.84	71.36	53.84	61.37
ECLM(trigram)	58.24	69.82	58.24	63.50
ECLM(+WSME-LM)	58.80	70.49	58.80	64.12
Text	91.16	92.17	91.16	91.66

Table 3: SLU evaluation result of TeleBank-Data

	Acc	Prec	Rec	F1
baseline (trigram)	78.54	84.23	78.54	81.28
ECLM(trigram)	81.16	85.57	81.16	83.31
ECLM(+WSME-LM)	83.60	87.72	83.60	85.61
Text	92.05	93.55	92.05	92.79

In table 2 and 3, we present the experimental results of the SLU task for baseline system and ECLM adaptation results. Accuracy (Acc), precision ($Prec$), recall (Rec), and F1-measure ($F1$) are shown on baseline, ECLM adaptations and text input. The last row of the table is the result of text input which assumes WER is 0%. Frame extraction performance is normally decreasing linearly by errors of speech recognizer. But we achieved improvement of frame extraction performance through ECLM adaptation. In comparison of CU-Data, SLU result of TeleBank-Data is better, because the speech recognition accuracy of TeleBank-Data is higher than that of CU-Data, and the definition of semantic frame slot is not much complex in TeleBank-Data.

5. Conclusion

The main issue on practical spoken language applications to provide interface between human and machine is how to overcome incompleteness of speech recognition and how to guarantee the reasonable end-performance of spoken language applications. Therefore, handling erroneously recognized output is the key of developing robust spoken language systems.

To address this problem, we proposed an ECLM adaptation to combine both domain environment characteristics and high-level linguistic knowledge. The ECLM technique is more adequate to deal with the erroneous outputs of ASR and can be adapted by task-specific text corpora and recognition results to improve the accuracy of speech recognition systems. The ECLM can be combined with general statistical language modeling methods, which successfully integrate linguistic information. In our ECLM adaptation framework, the WSME-LM is used as a re-scoring language model for error corrected candidate sentences. Moreover, the ECLM design allows for improvement via more sophisticated feature selection algorithm.

6. Acknowledgements

The authors thank Yulan He for providing the manually corrected semantic frame data in CU-Communicator Travel Data.

This research was supported from the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy.

7. References

[1] Bellegarda, J.R., Statistical language model adaptation: review and perspectives, Speech Communication, Volume 42, Issue 1, Pages 93-108, January 2004.

[2] Brown, P.F., Pietra, V.J.D., Pietra, S.A.D. and Mercer, R.L., The mathematics of statistical machine translation: parameter estimation, Computational Linguistics. 19(2), p263-311, 1993.

[3] Kneser, R. and Ney, H., Improved backing-off for m-gram language modeling, In Proc. of ICASSP, volume I, pages 181-184, Detroit, Michigan, May 1995.

[4] Lafferty, J., McCallum, A. and Pereira, F., Conditional Random Fields: Pobabilistic Models for Segmenting and Labeling Sequence Data, In Proc. of ICML, 2001.

[5] Pallett, D., Fisher, W., and Fiscus, J. Tools for the analysis of benchmark speech recognition tests, In Proceedings of the ICASSP, vol. 1. 97–100, Alburquerque, NM., 1990.

[6] Ringger, E.K. and Allen, J.F., Error correction via a post-processor for continuous speech recognition, In Proc. of ICASSP, 1996.

[7] Rosenfeld, R., Chen, S.F. and Zhu, X., Whole-Sentence Exponential Language Models: a Vehicle for Linguistic-Statistical Integration, Computer Speech & Language, 15(1), 2001.