Speech Recognition Error Correction Using Maximum Entropy Language Model

Minwoo Jeong, Sangkeun Jung, Gary Geunbae Lee

Department of Computer Science and Engineering, Pohang University of Science & Technology (POSTECH), San 31, Hyoja-Dong, Pohang, 790-784, Republic of Korea
{stardust, hugman, gblee}@postech.ac.kr

Abstract

A speech interface is often required in many application environments, such as telephone-based information retrieval, car navigation systems, and user-friendly interfaces, but the low speech recognition rate makes it difficult to extend its application to new fields. We propose a domain adaptation technique via error correction with a maximum entropy language model, which is a general and elegant framework to combine higher level linguistic knowledge. Our approach has the ability to correct both semantic and lexical errors in 1-best output from the black-box style speech recognizer, and can improve the performance of speech recognition and application system. Through extensive experiments using a speech-driven in-vehicle telematics information retrieval and spoken language understanding, we demonstrate the superior performance of our approach and some advantages over previous lexical-oriented error correction approaches.

1. Introduction

New application environments such as telephone-based retrieval, car navigation systems, and mobile information retrieval, often require a speech interface to conveniently process user queries. In these environments, keyboard input is inconvenient and sometimes impossible because of spatial limitations on mobile devices or instability in manipulating the devices.

However, because of the low recognition rate in current speech recognition systems, the final performance of speech applications such as speech-driven information retrieval (IR), question answering (QA), and speech dialogue systems is very low. The performance of the serially connected spoken QA system, based on the QA system from text input with a 76% performance and the output of the ASR, which operates at a 30% WER, was reduced to only 7% [1]. Harabagiu et al. [1] exposes several fundamental flaws in this simple combination of an automatic speech recognition (ASR) and QA system, including the importance of named entity information, and the inadequacies of current speech recognition technology based on n-gram language models.

The major problems with speech-driven IR and QA systems are decreased performance due to recognition errors in ASR systems. Erroneously recognized spoken queries often drop the precision and recall of IR and QA systems. For this reason, some appropriate adaptation techniques are required for overcoming speech recognition errors, such as post error correction techniques. ASR error correction can be a domain adaptation technique to improve recognition accuracy, and the primary advantage of the post error correction approach is its independence of the specific speech recognizer. Even if a speech recognizer can be regarded as a black-box, we can perform robust and flexible domain adaptation through the post error correction process.

One approach is based on a statistical method utilizing the probabilistic information of words in a spoken dialogue situation and the language models adapted to the application domain [2]. Ringer and Allen [2] first applied the noisy channel model to the correction of the errors in speech recognition. They tried to construct a general post-processor that can correct errors generated by any speech recognizer. The model consists of two parts: a channel model, which accounts for errors made by the ASR, and a language model, which accounts for the likelihood of a sequence of words being uttered. Also, they added the fertility model to handle split or merged errors such as 1-to-2 or 2-to-1 mapping errors. But this fertility model only slightly increased the accuracy in experiments [2], and we think the major reason is due to the data-sparse ness problem. Because substitution probability is based on the whole word-level, this fertility model requires enormous training data. Moreover, word-based simple n-gram language models are not suitable to capture syntactic and semantic errors.

In this regard, we propose a more improved model for post speech recognition error correction, which can be augmented with previous lexical-based approaches. In our approach, in addition to the lexical information, we use higher level knowledge sources in a speech transcription via maximum entropy language model (MELM).

2. Noisy Channel Error Correction Model

The noisy channel framework has been applied to a wide range of problems, such as speech recognition, statistical machine translation, and ASR error correction. The key idea of the noisy channel model is that we can model some channel properties by estimating the posterior probabilities.

The problem of ASR error correction can be stated in this model as follows [2]: For an input sentence, $O =$
\[ \hat{W} = \arg \max_{W} P(W|O) = \arg \max_{W} P(O|W)P(W) \quad (1) \]

At this point, we have a noisy channel model for ASR error correction, with two components, the language model \( P(W) \) and the channel model \( P(O|W) \). We call the model a word-based channel model because this model is based on word-to-word transformation. The word-based model focuses on inter-word substitutions, so it requires enough results of ASR and transcription pairs as a training data. Considering the cost of building a sufficient amount of correction pairs, we need a smaller unit than a word for overcoming data-sparseness.

For dealing with intra-word transformation, we propose a syllable-based channel model to deal with syllable-to-syllable transformation. This model is especially reasonable for Korean. In agglutinative languages such as Korean, a syllable is a basic unit of a written form like a Chinese character. For a syllable-based model, suppose \( S = s_1, s_2, \ldots, s_n \) is a syllable sequence of ASR output and \( W = w_1, w_2, \ldots, w_m \) is a source word sequence, then our purpose is to find the best word sequence \( W \) as follows:

\[ \hat{W} = \arg \max_{W} P(W|S) \quad (2) \]

We can apply the same Bayes’ rule and decompose the syllable-to-word channel model into a syllable-to-syllable channel model.

\[
P(w|s) = \frac{P(s|w)P(w)}{P(s)} \propto P(s|w)P(w)
\]

\[
\approx P(s|x)P(x|w)P(w) \quad (3)
\]

So, the final formula can be written as:

\[ \hat{W} = \arg \max_{W} (P(S|X)P(X|W)P(W)) \quad (4) \]

Here, \( P(S|X) \) is the probability of a syllable-to-syllable substitution, where \( X = x_1, x_2, \ldots, x_n \) is a source syllable sequence. And \( P(X|W) \) is a word model, which can convert the syllable lattice into a word lattice. The conversion can be done efficiently by a dictionary look-up. The last component \( P(W) \) is the language model, whose distribution can be defined using simple \( n \)-gram, or any other statistical language model.

To train the syllable-based channel model, we need a training data consisting of \( \{X, S\} \) pairs which are manually transcribed strings and ASR outputs. Also, we align the pair based on minimizing the edit distance between \( x_i \) and \( s_i \) by dynamic programming. Then, we applied the fertility to our syllable-to-syllable channel model. We set the maximum 2-fertility of the syllable, which was determined experimentally. Figure 1 shows an alignment for the syllable-model. In figure 1, user’s utterance, ‘한남대학교 건너 잡실로 가려면? (How do we go to Jamsil through Hanam bridge?)’, is actually recognized including some errors. For this erroneous recognition result, we can make transformation pairs for our syllable-based channel model. For example, (잡실로, 잡시 으로) pair can be divided into (잡, 잡), (실, 시), and (로, 으로) with fertility 2.

To apply the model, we should find the best probable word sequence in a given syllable sequence \( S \). This will be to return an N-best list of candidates according to the models, and then re-score these candidates by taking into account the language model probabilities. To re-score the candidates, we used the Viterbi search algorithm to find the best sequence. The system can generate a candidate word sequence network using our syllable channel model and a lexicon. Next, we can find the optimal sequence which has the best probability through Viterbi decoding by including a language model.

The language model \( P(W) \) is a crucial component in both speech recognition and speech recognition error correction. We used two kinds of language models: a word \( n \)-gram model and MELM. At the first level, a word \( n \)-gram model was used for capturing local dependencies and for fast processing. The MELM 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 error correction.

### 3. Error-Corrective Re-scoring Using MELM

The maximum entropy (ME) approach to language modeling was first successfully applied in [3]. One of the strength of the ME paradigm is the ability to incorporate arbitrary knowledge sources while avoiding fragmentation. So, the ME-based language models can combine \( n \)-gram features and other higher level linguistic knowledge in one unified framework [3, 4]. For this reason, we used MELM to deal with syntactic and semantic-level errors in our error correction model.

Using the chain rule, the language model \( P(W) \) can be decomposed as follows:

\[ P(W) = P(w_1 \ldots w_n) = \prod_i P(w_i|h_i) \quad (5) \]

where \( w_i \) is \( i^{th} \) word in the word sequence (sentence) \( W \), and \( h_i = \{w_1, \ldots, w_{i-1}\} \) is called the history of \( w_i \). In the ME framework, the conditional probability \( P(w|h) \) can be
The Gaussian prior is a powerful tool for smoothing general exponential models, and like other ML methods is prone to overfitting of the training corpus. If the constraints \( \lambda_i \) are consistent, there exists a unique solution called a minimum discrimination information (MDI) solution or ME solution when the prior is uniform. This unique ME/MDI solution can be found by several iterative methods such as generalized iterative scaling (GIS), improved iterative scaling (IIS) or the limited-memory variable metric method, which is a limited-memory version of the quasi-newton method (also called L-BFGS). The L-BFGS was the most effective training method for the ME model [5]. We used the L-BFGS method to estimate the parameters of MELM for our error correction model.

The MELM can be viewed as an ML training for exponential models, and like other ML methods is prone to overfitting of the training data [6]. Moreover, we need a technique to smooth our MELM, because unseen events in the training data have zero probability. We adopt a Gaussian prior smoothing from several proposed methods for ME models. The Gaussian prior is a powerful tool for smoothing general ME models, and can work well in the language models [6].

### 4. Experiments

We performed several experiments on the domain of in-vehicle telematics IR related to navigation question answering services. The speech transcripts used in the experiments were composed of 462 queries, which were collected by 1 male speaker in a real application situation. We also used two Korean speech recognizers: a speech recognizer made by LG-Elite (LG Electronics Institute of Technology) and a Korean commercial speech recognizer, ByVoice (refer to http://www.voicetech.co.kr). In addition, we divided the 462 queries into 6 different sets, and evaluated the results of 6-fold cross validation for each model.

We constructed a syllable-based channel model for error correction and integrated a MELM to re-score the N-best lists of the channer model error correction. For experiments, we used bi-gram language models generated by SRILM toolkit and various MELMs generated by MaxEnt toolkit. We used several features for MELM as in the following:

\[
\begin{align*}
\hat{P}(h) \cdot \prod_w P(w|h) \cdot f_i(h, w) &= \hat{E}_\lambda f_i(h, w) \tag{7}
\end{align*}
\]

Table 1 presents the experiment results of WER and SER of two types of baseline ASR systems, error correction using syllable-based channel model and combined models with MELMs for re-scoring error correction.

Table 1: WER and SER of baseline ASR, noisy channel error correction (NC-ECO), and several MELM models with different feature sets. WER (SER) each designates the errors for each two different ASRs used.

<table>
<thead>
<tr>
<th></th>
<th>WER_1</th>
<th>WER_2</th>
<th>SER_1</th>
<th>SER_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>39.0%</td>
<td>34.8%</td>
<td>12.4%</td>
<td>16.4%</td>
</tr>
<tr>
<td>NC-ECO</td>
<td>27.0%</td>
<td>30.8%</td>
<td>10.3%</td>
<td>15.4%</td>
</tr>
<tr>
<td>+MELM(D)</td>
<td>23.6%</td>
<td>25.3%</td>
<td>9.1%</td>
<td>13.6%</td>
</tr>
<tr>
<td>+MELM(D+T)</td>
<td>23.2%</td>
<td>24.7%</td>
<td>9.0%</td>
<td>13.2%</td>
</tr>
<tr>
<td>+MELM(D+S)</td>
<td>22.9%</td>
<td>24.8%</td>
<td>8.7%</td>
<td>13.2%</td>
</tr>
<tr>
<td>+MELM(ALL)</td>
<td>22.6%</td>
<td>24.6%</td>
<td>8.6%</td>
<td>13.0%</td>
</tr>
</tbody>
</table>

- **Distance-2 n-gram**
  Distance-2 n-gram (d2-n-gram) is an n-gram with 2 words back to the word to be predicted. For example, d2-bigram predicts \( w_i \) based on \( w_{i-2} \) and d2-trigram predicts \( w_i \) based on \( w_{i-3}, w_{i-2} \). Conventional n-gram features are not used in our MELM because these features are already incorporated in the previous error correction model.

- **Trigger pairs**
  A trigger pair, which is a pair of highly correlated words in the same document (or in a sentence in our experiment data), has been successfully used in MELM [3] as a vehicle to treat long-distance dependencies. We used Adam Berger’s trigger toolkit to select trigger pairs from training data.

- **Syntactic features**
  Some syntactic features can be included in MELM. We used two types of features: part-of-speech tag and chunk.

To measure error correction performance, we use word error rate (WER) and syllable error rate (SER):

\[
\begin{align*}
\text{WER} &= \frac{|S_w| + |I_w| + |D_w|}{|W_{truth}|} \tag{8}
\end{align*}
\]

\[
\begin{align*}
\text{SER} &= \frac{|S_s| + |I_s| + |D_s|}{|S_{truth}|} \tag{9}
\end{align*}
\]

\( |W_{truth}| \) is the number of original words, and \( |S_{truth}| \) is the number of original syllables (or characters in Chinese and Japanese). WER and SER are rates of errors, which include substitution (|S|), insertion (|I|), and deletion (|D|) errors of each unit.

1Available at <http://www.speech.sri.com/projects/srilm/>
2Available at <http://www.nlplab.cn/zhangle/>
Table 2: CER of text input (perfect speech recognition; WER zero), baseline ASR (WER 39% and 35%), noisy channel error correction (NC-ECO), and several MELM models with different feature sets.

<table>
<thead>
<tr>
<th></th>
<th>CER.1</th>
<th>CER.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Input</td>
<td>20.5%</td>
<td>20.5%</td>
</tr>
<tr>
<td>Baseline</td>
<td>33.3%</td>
<td>31.9%</td>
</tr>
<tr>
<td>NC-ECO</td>
<td>28.1%</td>
<td>31.1%</td>
</tr>
<tr>
<td>+MELM(D)</td>
<td>27.7%</td>
<td>30.1%</td>
</tr>
<tr>
<td>+MELM(D+T)</td>
<td>27.8%</td>
<td>29.6%</td>
</tr>
<tr>
<td>+MELM(D+S)</td>
<td>27.6%</td>
<td>29.6%</td>
</tr>
<tr>
<td>+MELM(ALL)</td>
<td>27.5%</td>
<td>29.4%</td>
</tr>
</tbody>
</table>

1. MELM(D) denotes MELM trained with d2-n-gram features; MELM(D+T) was trained with d2-n-gram and triggers; MELM(S) was trained with d2-n-gram and syntactic features; MELM(ALL) was trained with all the features.

The WER of each baseline system was about 39% and 35% on the utterances in in-vehicle telematics IR domain. This result shows that the error correction of speech recognition result is a viable approach to improve the performance. Using both baseline ASR systems, we achieved a 42% and 29% of error reduction rate for WER.

To validate our approach in another application, we conducted an experiment on spoken language understanding (SLU) system. In our SLU experiments, the SLU system can produce concept sequences, which can be defined by extracted frames by a slot/filling technique. To measure the performance of SLU, we use the concept error rate (CER):

\[ CER = \frac{|S_c| + |I_c| + |D_c|}{|C_{truth}|} \] (10)

|C_{truth}| is the number of manually extracted concepts, and CER means the similarly defined error rate of misunderstood concepts in SLU.

In table 2, we present the experimental results of SLU in text input, baseline systems and each of our error correction model. When the CER of text input (WER zero) are about 20.5%, the SLU performance of spoken inputs are decreased by 13% and 11% in this domain (baseline). However, adaptation via error correction can increase the performance of SLU systems as shown in the table 2. This result shows that the speech recognition error correction is a viable approach to improve the final SLU performance, in addition to the ASR performance.

5. Conclusion

We proposed an improved noisy channel model and error corrective re-scoring model with combined high-level linguistic knowledge using MELM for speech recognition error correction, and demonstrate the superior performance in both ASR and domain-specific SLU applications.

We think that the syllable-based channel model can improve the performance of black-box style speech recognition systems and capture more specific error patterns. Because the ME framework is elegant and general, MELM can be a theoretically elegant model and integrate successfully higher-level linguistic knowledge. In our error correction model, MELM is used as a re-scoring language model for error corrected candidate word sequences. However, MELM has more chance to improve the performance via more sophisticated feature induction or selection algorithms.

Training a MELM is computationally challenging, and sometimes infeasible. So, we are working on some techniques to overcome the slow training time of MELM [4, 8].

6. Acknowledgements

This work was partly supported by Jungki KeoJeom Project (MOCIE, ITEP) and by 21C Frontier Project (MOST).

7. References


