Multimodal Interface for Error Correction in Speech Recognition

Koichi Shinoda
Tokyo Institute of Technology
shinoda@cs.titech.ac.jp
http://www.ks.cs.titech.ac.jp/english/yuan.html

1. Project Goal

The goal of our project is to design an easy-to-use error correction interface for mobile speech-to–text applications. In most error correction interfaces, when a user finds an error word in the recognition result, he/she firstly marks it and then either selects the correct word from a candidates list provided by the interface, or inputs the correct word via a single input modal such as speech, pen or keyboard. The error correction process is cumbersome and time-consuming. In this user-machine interaction process, there will be a lot of useful, correct information generated, which has not been used in the process of error correction before. Our goal is utilize such rich information to realize efficient error correction interface.

To achieve this goal, we propose a multi-modal interface (Fig. 1). The error correction process is divided into the following steps: 1) User corrects errors from left to right. He/she locates an error region using pen gesture; 2) System runs a re-ranking process specified for the error region, and shows a candidates list for the correct word; 3) If the correct word is in the list, User choose it and stop; otherwise, he/she starts inputting the correct word either by speech, keyboard and pen, or by their combination. System updates the candidates list at each user input until the correct word appears in the list.

The following two problems should be solved: 1) how to use the information obtained in the error correction process to generate a more accurate candidates list; 2) how to repair the errors which occur due to the out-of-vocabulary (OOV) errors.

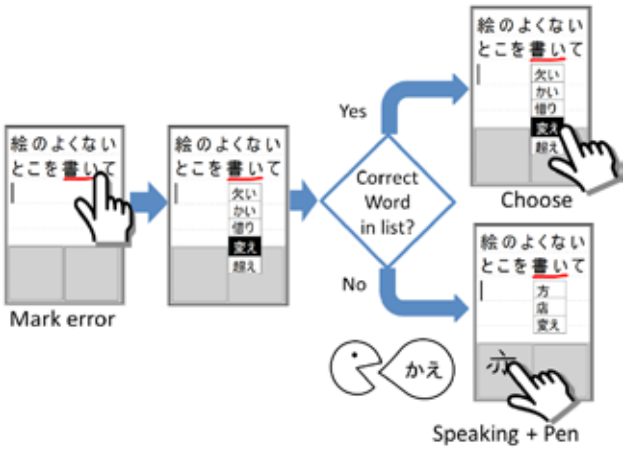

Figure 1: Multi-modal Error Correction Interface

2. Technical breakthrough

2.1 How to generate a more accurate candidates list

First we assume that users correct errors from left to right. Our interface waits before starting the error correction process until the time when the user finishes marking the other error word in the succeeding word sequences. On this assumption we can safely assume that all the words preceding the error region are correct or already been corrected, and that the words succeeding the error region and preceding the next error region are all correct.

We proposed a new n-gram language model (LM) for error correction, in which the occurrence probability of a word depends not only on preceding words but also on succeeding words. We name it Bidirectional Language Model (BLM). Different from the standard LM, it uses the future information to predict the word, which has never been used before. In this project period, we deal with the case only a sentence involves only one error word. Further, we deal with a Trigram BLM, where the occurrence probability of a word depends on the one preceding word and one succeeding word. We make a candidate for the user

correction as follows. First we apply a standard Trigram LM to do decoding, create a word lattice. Then we delete hypotheses in which the proceeding words or the following one word are incorrect, and reformulate the word lattice using our proposed Trigram BLM. Finally we transform the word lattice to a confusion network.

We evaluated our method on the Corpus of Spontaneous Japanese (CSJ). 953 academic lectures having 228 hours length were used for training acoustic HMMs. The transcribed texts of 2496 were used for LM training. We extracted 1067 sentences, each of which has one error word. In this experiment, we use an n-best list instead of a word lattice. Table 1 compares the perplexity of several methods. The perplexity of BLM was significantly better than that of the standard LM. Table 2 shows BLM improves the rank of the correct word in the candidates list.

In future first we implement our BLM to word lattice rescoring. We will extend our method to the case when multiple errors exist in one sentence. There may be the other assumptions that can be used for an error correction interface. We will implement a new LM on such assumptions.

Bigram	Trigram	4gram	Trigram BLM
93.8	77.0	75.7	36.9

Table 1: Perplexity

Higher	Equal	Lower
239	641	187

Table 2: The number of correct words whose ranks becomes higher/equal/lower using Trigram BLM compared with using Standard Trigram LM after re-ranking

2.2 How to solve OOV problems in multimodal interface

We develop a multimodal interface for OOV problems, in which User simultaneously input speech and pen-gesture. This interface utilizes the output information from search engine (e.g. rank information) and the result from an instant very-large vocabulary speech recognition decoder specified for the error region to generate efficient candidates list. T3 decoder (Tokyo Tech Transducer-based Speech decoder) is used as the decoder.

3. Innovative Applications

The product of this project can be widely used as a speech interface for mobile devices such as PDAs and smartphones. Using our proposed interface, users can find the correct word in the candidates list more often, which will ease the burden on the user in the error correction process. Even if the user cannot find the correct word in the candidates list, the system supports other input modalities, speech and pen input. Compared to unimodal input, the correction accuracy, correction speed and usability will be improved.

4. Academic Achievement

Yuan Liang, Koichi Shinoda, and Sadaoki Furui, " Language Model for Efficient Error Correction in Speech Recognition " , Proc. of 2012 Spring Meeting of Acoustical Society of Japan, Yokohama, March (2012)

5. Achievement in Talent Fostering

A graduate student, Ms. Yuan Liang, is involved in this project. She has learned speech technology, human interface, and software development in our lab.

6. Project Development

The project is will be supported by the Grant-in-Aid for Scientific Research from MEXT.