Modeling the Differences between Spoken Language and Written Language for Automatic Speech Transcription

Tatsuya Kawahara, Kyoto University

Natural language has been long used for conveying information and knowledge in the human society. It has two forms: spoken language for real-time communication and written language for off-line communication. As the recent innovation of ICT (Information and Communication Technology) has fluidized the role of these two forms, automatic transformation of the spoken language and written language is also getting an important research topic. Specifically, as digital recording and archiving of lectures and meetings has become pervasive, we address the problem of automatic transcription, or automatic speech recognition (ASR), of these kinds of real-world human-to-human communication.

We have compiled two large-scale corpora which contain aligned sets of audio files, their faithful transcripts, and their document-style texts. One is a collection of records of oral presentations given at technical conferences, and the other is a set of meeting records of the National Diet (Congress) of Japan. It is observed that more than 10% of words of the faithful transcripts of the utterances were corrected or deleted in making the cleaned texts. This editing process is modeled as a statistical transformation process, which will constitute an “intelligent” transcription system. By inversing the process, we can also model a process of giving a speech from the document-style text, such as talking based on a pre-print technical paper. This will help enhance the language model for ASR.

We have also studied pronunciation variation from the orthodox base forms, observed in spontaneous speech. There are frequent patterns on reduction/deletion of phones depending on their phonetic context. This is also formulated as another statistical transformation model and used for an elaborate pronunciation model for ASR.

Based on these models, we are developing two intelligent transcription systems. One is a meeting transcription system for the House of Representatives, which will assist stenographers to make official minutes of the meetings. The other is a lecture captioning system, oriented for real-time note-taking for deaf students. In both systems, effects of the language model and the pronunciation model are confirmed, and automatic cleaning of the transcripts is useful for improving readability. We are now investigating the feasibility and further requirements of our technology in the real field.

Keywords: Speech, Language, Automatic Speech Recognition, Natural Language Processing