Public Abstract
Xiaojia Zhang
MS
Language modeling for automatic speech recognition in telehealth
Advisor: Dr. Yunxin Zhao

Graduation Term: FS 2005

Telehealth system is becoming an advanced way of combining telecommunication and information technology to provide health care to elderly people or people living in rural areas with high efficiency and quality. The project of automatic speech recognition for telehealth, carried out in the Spoken Language and Information Processing Laboratory in the Department of Computer Science, University of Missouri-Columbia, captions doctors' speech for patients with hearing disabilities to help them better understand doctors' questions and explanations. Language modeling which is an important part of automatic speech recognition faces challenges in this specific task, since doctors' speech contains a large amount of infrequently used medical terms in a spontaneous conversational style, for which there are insufficient in-domain data for model training.

In order to train a robust language model that is important in building a reliable automatic speech recognition system, research work was done to investigate methods for improving model performance, and the methods and outcomes are discussed in this thesis. Various datasets were collected, analyzed and used as supplementary resources to the insufficient in-domain data. By specifically treating datasets from different data resources based on their styles and topics, and by training class and word trigram language models from in-domain and out-of-domain datasets separately, we developed an intuitive but efficient modeling method that takes advantages of both word and class language models. Training class language model based on semantic categorization of medical terms was seen to be more meaningful and practical than some commonly used word clustering methods for this particular task. Language model derived from an optimized combination of word and class language models brought significant reduction on both perplexity and recognition word error rate.