LANGUAGE MODELING FOR AUTOMATIC SPEECH RECOGNITION IN TELEHEALTH

Xiaojia Zhang

Dr. Yunxin Zhao, Thesis Supervisor

ABSTRACT

Standard statistic n-gram language models play a critical and indispensable role in automatic speech recognition (ASR) applications. Though helpful to ASR, it suffers from a practical problem when lacking sufficient in-domain training data that come from same or similar sources as the task text. In order to improve language model performance, various datasets need to be used to supplement the in-domain training data. This thesis investigates effective approaches to language modeling for telehealth which consists of doctor-patient conversation speech in medical specialty domain. Efforts were made to collect and analyze various datasets for training as well as to find a method for modeling target language. By effectively defining word classes, and by combining class and word trigram language models trained separately from in-domain and out-of-domain datasets, large improvements were achieved in perplexity reduction over a baseline word trigram language model that simply interpolates word trigram models trained from different data sources.