Language Modeling for Speech Recognition of Spoken Cantonese

In this seminar, language modeling for Cantonese speech recognition are discussed. Cantonese is a spoken dialect of Chinese language which shows lexical and grammatical differences from standard Chinese. As a spoken dialect, Cantonese is not used in formal written documents and published materials, making the collection of written Cantonese texts for language modeling difficult. We propose to solve this problem by translating standard Chinese texts, which are much easier to find, into written Cantonese by a rule-based method. The rules are obtained from a small standard Chinese-written Cantonese parallel corpus automatically by the transformation-based learning. Language models for Cantonese are trained from a large newspaper text corpus which are originally in standard Chinese but is translated by the rules. Experimental results confirm that the usage of standard Chinese language model on Cantonese speech recognition task would degrade the accuracy significantly, even if the speech is spoken in a more formal manner. The rule-based method can improve the recognition results of formal Cantonese broadcast news task by absolute 4.2% and colloquial Cantonese task by absolute 6%. The improvement can be more significant if a better acoustic model is used. While the rule-based method can achieve a reasonably good performance on the formal Cantonese task, the accuracy on the colloquial Cantonese task is relatively lower. The performance of the colloquial Cantonese task can be improved by language model adaptation with a limited amount of colloquial Cantonese texts.