Session 2: Language Modeling

Xuedong Huang, Chair

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

This session presented four interesting papers on statistical language modeling aimed for improved large-vocabulary speech recognition. The basic problem in language modeling is to derive accurate underlying representations from a large amount of training data, which shares the same fundamental problem as acoustic modeling. As demonstrated in this session, many techniques used for acoustic modeling can be well extended to deal with problems in language modeling or vice versa. One of the most important issues is how to make effective use of training data to characterize and exploit regularities in natural languages. This is the common theme of four papers presented here.

In the first paper, Ronald Rosenfeld from CMU described his thesis work on maximum entropy language modeling. The maximum entropy model is able to incorporate multiple constraints consistently. Although the maximum entropy model is computationally expensive, it could potentially help speech recognition significantly as the approach allows us to incorporate diverse linguistic phenomenon that can be described in terms of statistics of the text. With the maximum entropy approach, Ronald demonstrated that trigger-based language adaptation reduced the word error rate of CMU's Sphinx-II system by 10-14%.

Rukmini Iyer from BU then presented her recent work on mixture language modeling. The model is an m-component mixture of conventional trigram models, which are derived from clustered WSJ training data. As we know, the mixture acoustic model has significantly improved many state-of-the-art speech recognition systems. Rukmini demonstrated here that mixture language models also reduced the word error rate by 8% using the BU speech recognition system.

Ryosuke Isotani from ATR described a very interesting method that integrates local and global language constraints for improved Japanese speech recognition. The approach exploited the relationship of function words and content words, and used the combined language model for speech recognition. As a result, Ryosuke demonstrated that the word error rate of the proposed language model was comparable to that of the trigram language model, but the parameter size was significantly reduced. It would be interesting to see if the proposed model can be applied to different languages, and if it remains effective with a larger data base.

Finally, Rich Schwartz from BBN presented a paper that addresses three problems associated with language modeling. He first demonstrated that additional training data substantially improved the language model performance. Second, he introduced a method to minimize the difference between the language model training text and the way people speak. Third, he showed that by increasing the vocabulary size, the recognition accuracy did not degrade significantly. This somewhat alleviated problems associated with new words in the test material.