

Department of Applied Mathematics and Statistics
The Johns Hopkins University

SEMINAR

Fred Jelinek
Dept. of Electrical & Computer Engrg.
The Johns Hopkins University

April 5, 2007
304 Whitehead Hall
Refreshments: 3:30 p.m.
Seminar: 4:00 p.m.

LANGUAGE MODELING BY RANDOM FORESTS

ABSTRACT

Automatic Speech Recognition is based on several components: signal processor, acoustic model, language model, and search. In this talk, we explore the use of Random Forests (RFs) in language modeling, the problem of predicting the next word based on words already seen. The goal is to develop a new language model smoothing technique based on randomly grown Decision Trees (DTs). This new technique is complementary to many of the existing techniques dealing with data sparseness.

Random forests were studied by Breiman in the context of classification into a relatively small number of classes. We study their application to n -gram language modeling which could be thought of as classification into a very large number of classes. Unlike regular n -gram language models, RF language models have the potential to generalize well to unseen data, even when histories are long (> 4). We show that our RF language models are superior to regular n -gram language models in reducing both the entropy and the word error rate in a large vocabulary speech recognizer.