

COE Quarterly Technical Exchange

Tuesday, January 13, 2009, 1:00 – 5:00 pm

Stieff Building/810 Wyman Park Drive Directions: <http://web.jhu.edu/HLTCOE/contact.html>

This event is UNCLASSIFIED

Registration required: email Karen Daughton-Thomas at: kdaught2@jhu.edu

| Time | Topic / Speaker |
|-------------|---|
| 1:00 – 2:00 | Detailed Analysis of Self-Training Behavior Scott Novotney, Rich Schwartz |
| 2:00 – 3:00 | Statistical Validation: Delayed- Decisions and Noisy Hypothesis Testing for Machine Learning Chris White |
| 3:00 – 3:15 | Break |
| 3:15 – 4:15 | Speech Research Software Laboratory -- Overview and Demo Hugh Secker-Walker, Ken Basye |
| 4:15 – 4:30 | Closing |

Note: The next HLTCOE Quarterly Technical Exchange will take place in April and will focus on text topics.

NOTE: In the case of inclement weather, please go to the HLTCOE home page, www.hltcoe.org, to see if there is an announcement about postponement of this event. We will also be sending out an email to alert attendees if the weather prohibits holding this meeting.

Abstracts

Detailed Analysis of Self-Training Behavior

Scott Novotney, Rich Schwartz

Our work over the previous year established operating points for different dimensions of speech recognition: acoustic and language modeling. Now we demonstrate where self-training succeeds and under what resource conditions it provides the most benefit. First we demonstrate that self-training works successfully with tougher acoustic conditions using the Callhome data set. Then, we continue the tear down of available knowledge sources and constrain the vocabulary to that present in ten hours of speech. Finally, by digging beneath word error rate and analyzing individual word performance, we show that self-trained models learn new words. More importantly, the words that only appear in the large unlabeled audio corpus improve by a much larger amount than the words that are in training.

Statistical Validation: Delayed- decisions and Noisy Hypothesis Testing for Machine Learning

Chris White

We propose a framework and set of procedures for leveraging potentially unlimited amounts of unlabeled data to improve statistical models. We build upon results in hypothesis testing (Poor), sequential analysis (Wald), and robust statistics (Huber) in order to make principled system-level decisions based on whatever data (labeled or not) are available. A central aspect of these procedures is the application of a non-parametric hypothesis test about the significance between two possibly indistinguishable configurations. Initial results from pronunciation validation will also be presented.

Speech Research Software Laboratory -- Overview and Demo

Hugh Secker-Walker, Ken Basye

The TTO5 Research Software Laboratory is a toolkit designed to be a flexible and robust vehicle for experimentation. The toolkit consists of tools and policies that support a diverse set of high-level design and usability goals. A primary goal is to provide a library of modules which can be assembled easily to build new experiments; these modules and assemblies are "transparent" in the sense that intermediate results and internal state can be inspected easily which allows researchers to get useful insights quickly. Other goals include facilitating processing of unbounded streams of data and providing an environment that facilitates maintaining reproducibility of results. This talk gives an informal, high-level overview of the toolkit and a demonstration of some of the capabilities that enable rapid experimental development.