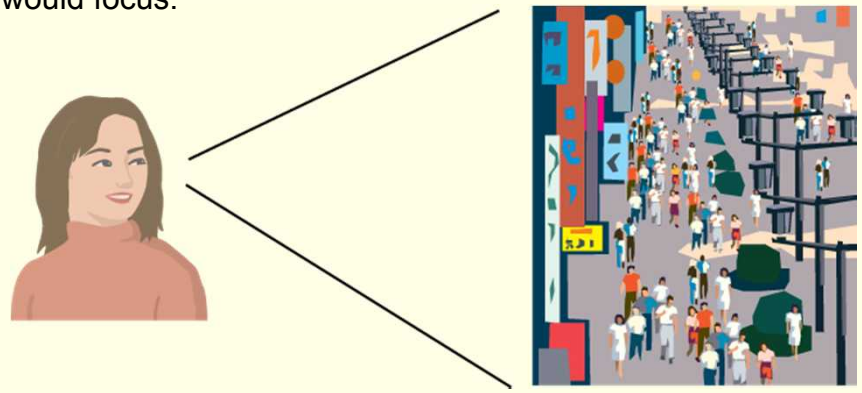
SAND2008-4368C

Detection of Surprise in Document Corpora through Human Modeling

Travis Bauer
Cognitive Systems Research and Development
tlbauer@sandia.gov
505-284-8723

The Video Research

In previous work, Laurent Itti and Pierre Baldi (both from the University of Southern California) showed a relationship between the underlying statistics of pixels in a video and where in the video a person would focus.



Previous research showed a correlation between where in a video a person watched and the underlying statistics of the pixels in the video.

Surprise is violation of expectation

- Build up expectations using some measure of the source material.
- Look for places where those expectations are violated.
- So surprise is effectively something that is both:
  - Relevant (we've seen enough to have built up expectations)
  - New (We are seeing something we haven't seen before)

"Surprise" is a violation of expectation. In the video work, these "expectations" were built up based on salience and novelty.

The Question

"Can you do that in text?"

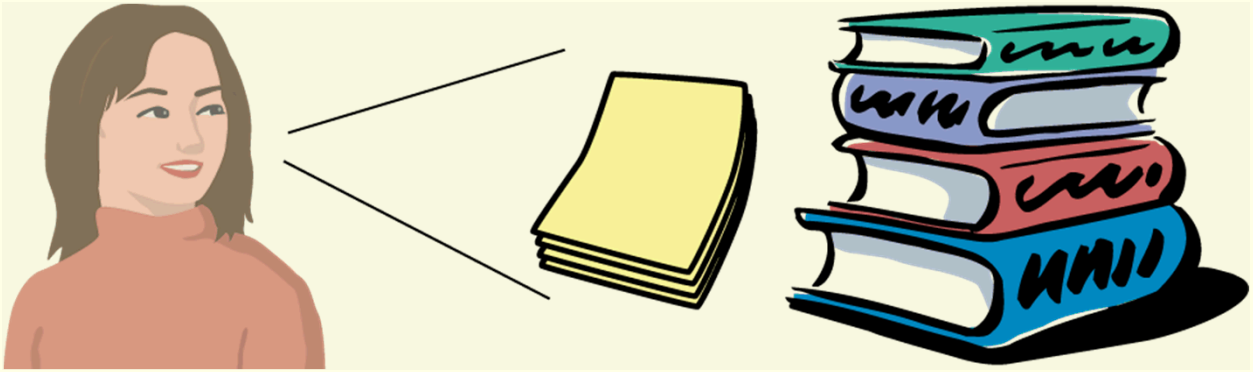
YES

The Basic Analyst Problem

Scenario: General desire from the user to find new "surprising" information in new documents that are read.

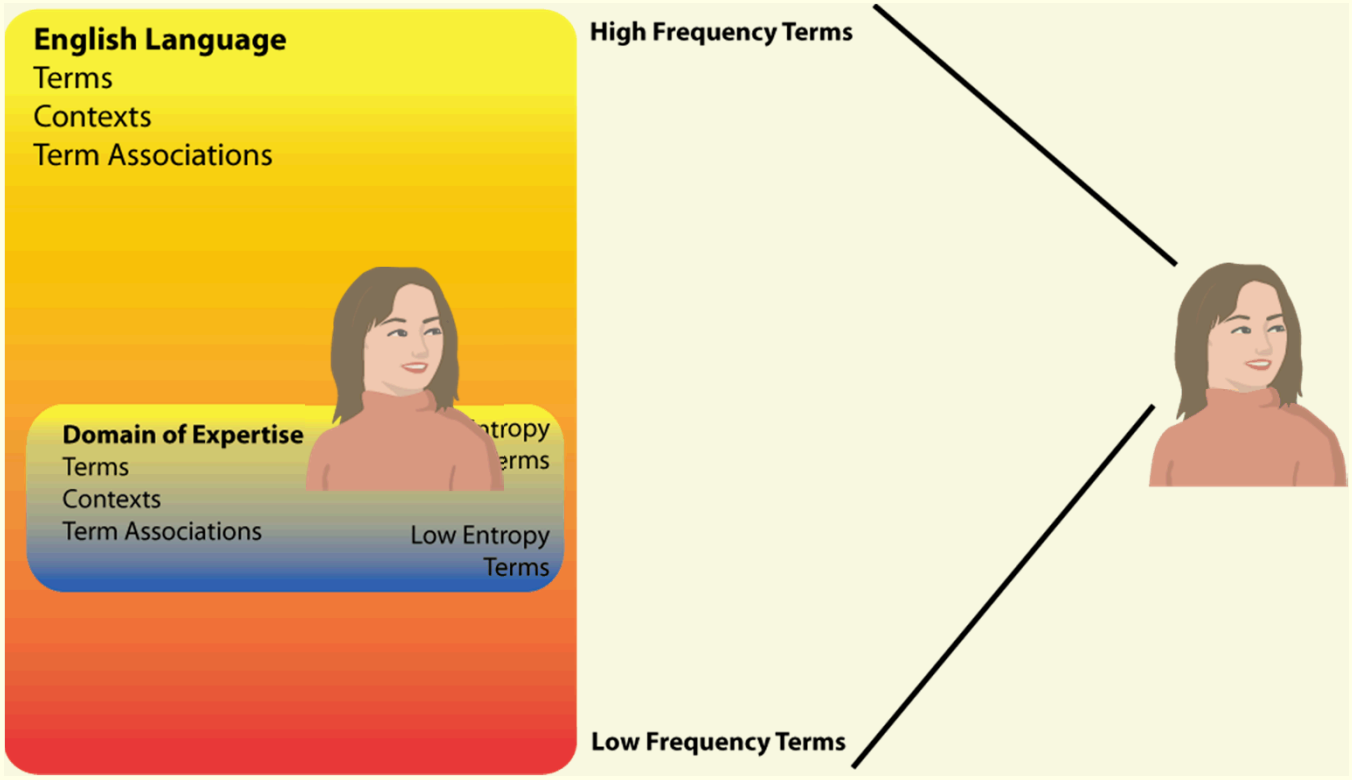
Basic statistical technique

Find what's common in an individual document × Find what's uncommon in the corpus = Vector



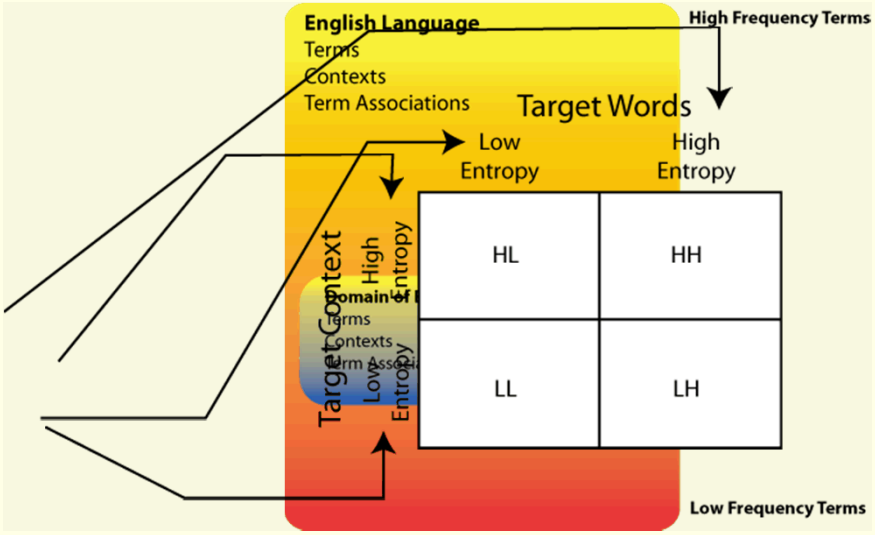
Approaches that rely on statistical analysis often rely on comparing the statistics of a corpus to term occurrences of an individual document as a basis for evaluating what the user would find interesting

We shouldn't ignore the user



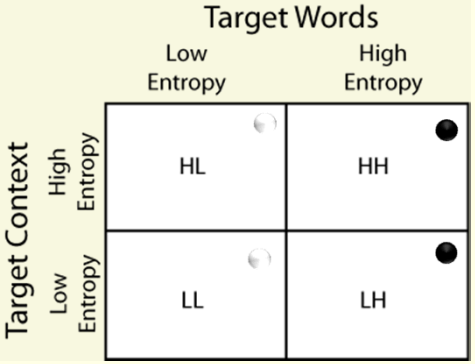
Users have complex models of their domains of expertise. Understanding that model will let us build more sophisticated data mining techniques for responding to information requests.

Experimental Setup

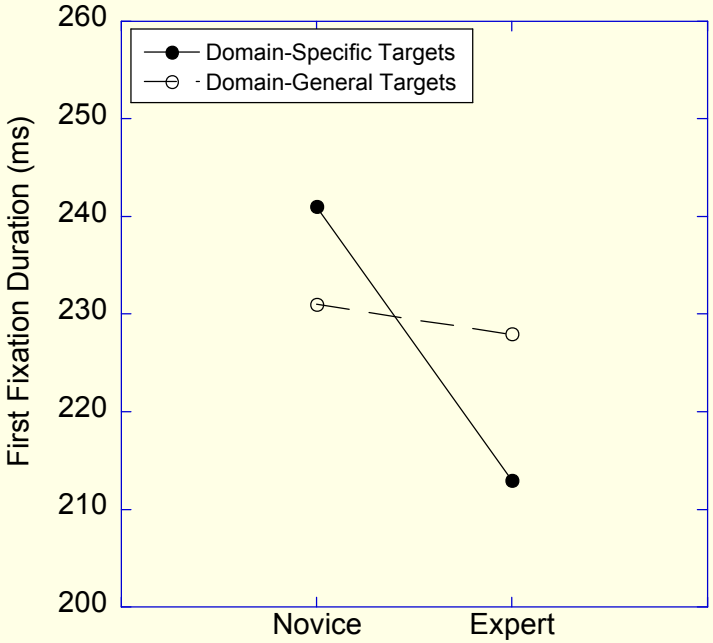


How Eye Trackers Get Used

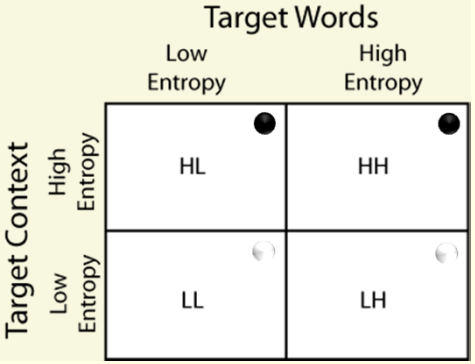
- First fixation duration
  - The amount of time the eye focuses on a term the first time it fixates on it
- Total gaze duration
  - The total amount of time the eye focuses on a term while reading
- Probability of skipping target words
  - How likely a term is to be passed over during reading
- Regressions
  - How often the person moves back to fixate on a term during reading



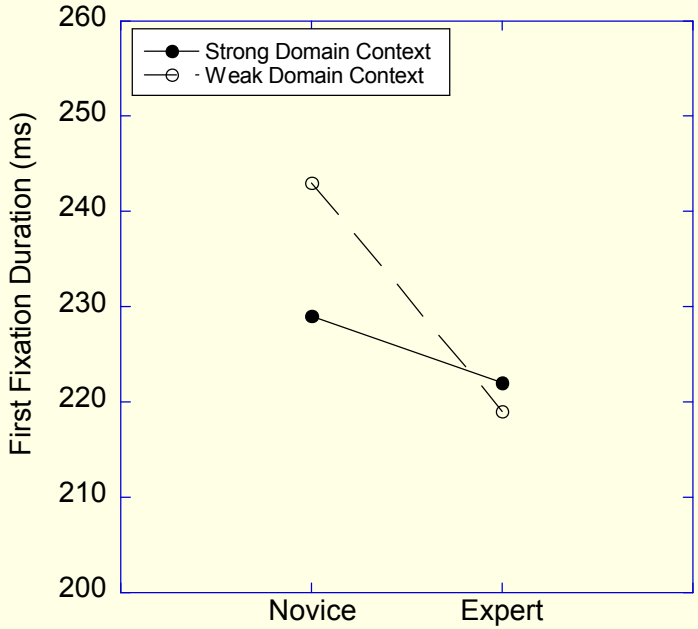
Basic Findings



The first fixation durations for novices and experts were more similar on domain general targets than on domain specific targets.



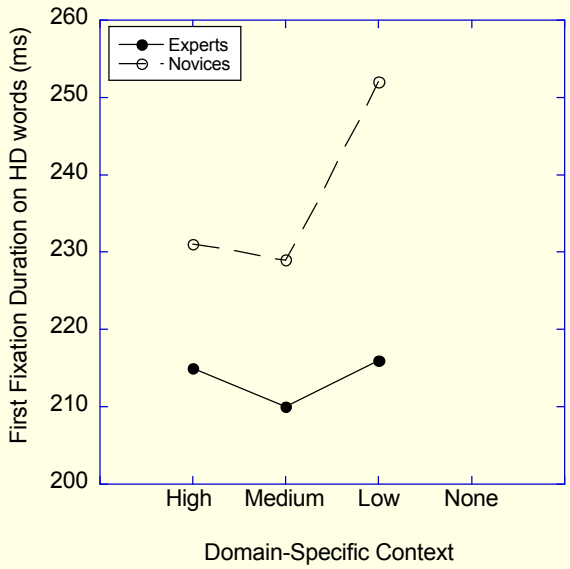
Basic Findings

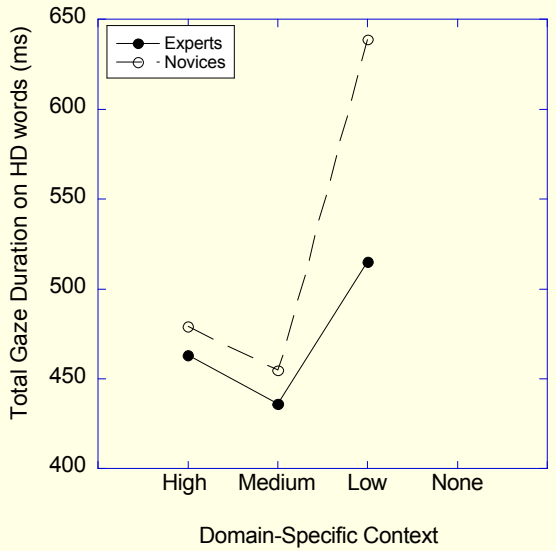


For novices, first fixation durations are quicker in sentences dominated by domain vocabulary. Experts do not show this affect.

Basic Findings

Target Context	Target Words: Low Entropy	Target Words: High Entropy
High Entropy	(medium) HL	(high) HH
Low Entropy	LL	(low) LH





Experts read domain specific vocabulary more quickly than novices, but the speed is affected by the context within which the domain specific vocabulary occurs.

The new things we know now that we didn't know before

- Eye movements are sensitive to the match between knowledge of the reader and the statistical properties of the terms in the text.
- People are encoding the entropy of terms in a domain and their eye movements are affected by it.
- Experts:
  - Read domain-related targets more quickly
  - Were more likely to be able to recall what they had read in an incidental memory task.
- Novices, by contrast,
  - Took longer to process the domain-specific concepts, but only when they were embedded in domain-general text that would afford them entry into meaning.
  - Were less likely to be able to recall what they had read in an incidental memory task.

Who Cares?
(or What can we do with this knowledge?)

Further Explore The Feature Space

Term Associations

Volume of Text Read

Context such as Project Phases

Fine Grained Studies of Ranges in the Feature Space

Different Levels of Entropy

Robustness

Differences between common English and Domain

Prototyping

Personal Assistants

Model Building Tools

Exploratory Algorithm Development

Development of better eye tracking hardware

- Better insight into how people process text and domain knowledge
- Algorithms for modeling individuals by watching their eye movements
- Technology to assist users by anticipating their needs
- Algorithms that model people based on text they produce or access
- Human behavior inspired models for text analysis
- Development of better test sets for testing information retrieval systems
- Eye tracker hardware that is robust for specific things needed for modeling

There is a clear path from these initial findings to useful applications

Questions? Surprises? Surprising Questions?

Travis Bauer
6343 Cognitive Systems Research and Development
tlbauer@sandia.gov
In collaboration with
Elizabeth Stine-Morrow and Matthew Shake
University of Illinois

Further Exploration of the Feature Space

PROGRAM DESCRIPTION

TECHNICAL DETAILS

TECHNICAL DETAILS

Studies of Ranges in the Feature Space

PROGRAM DESCRIPTION

TECHNICAL DETAILS

TECHNICAL DETAILS

Exploratory Algorithm Development

PROGRAM DESCRIPTION

TECHNICAL DETAILS

TECHNICAL DETAILS

Eye Tracker Development

PROGRAM DESCRIPTION

TECHNICAL DETAILS

TECHNICAL DETAILS