

GALE Formative Utility Lab

Lynne Plettenberg, Heather Halpin,
Allison Powell, Tim Clarke
JHU CLSP Summer School
June 30, 2006

Schedule

1:30- 2:00 Intro and training

2:00- 2:30 User Study Task

2:30 – 3:00 Discussion/Debrief

3:00- 3:30 Break

3:30- 4:00 GALE Challenges: ASR, MT, Segmentation

4:00- 4:30 Development Task (in small groups)

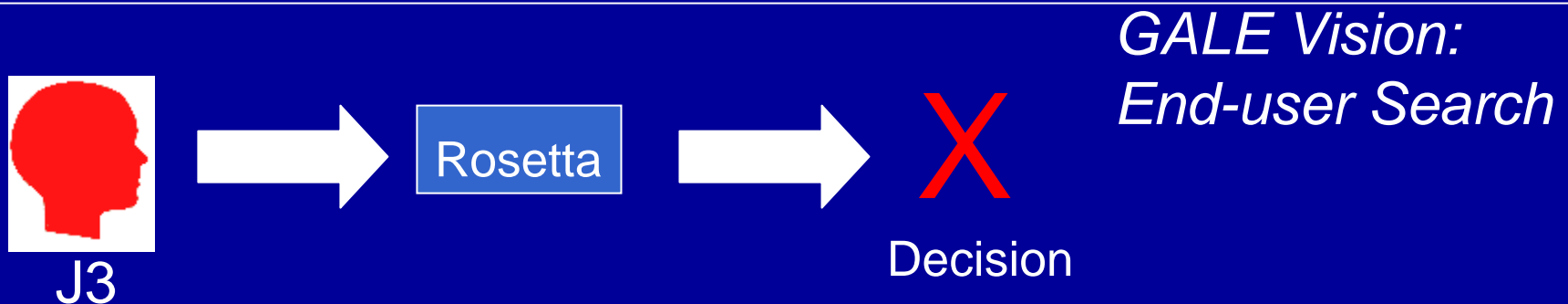
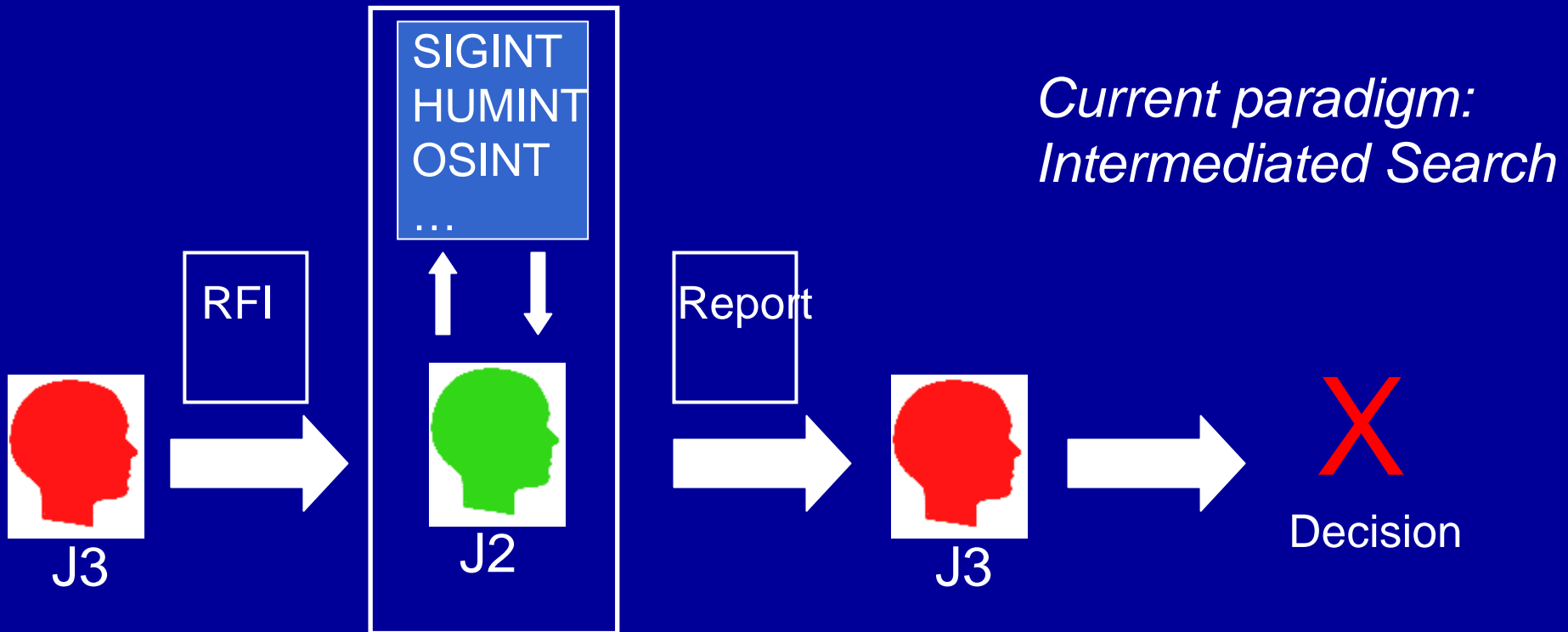
4:30- 5:00 Discussion/Debrief

GALE

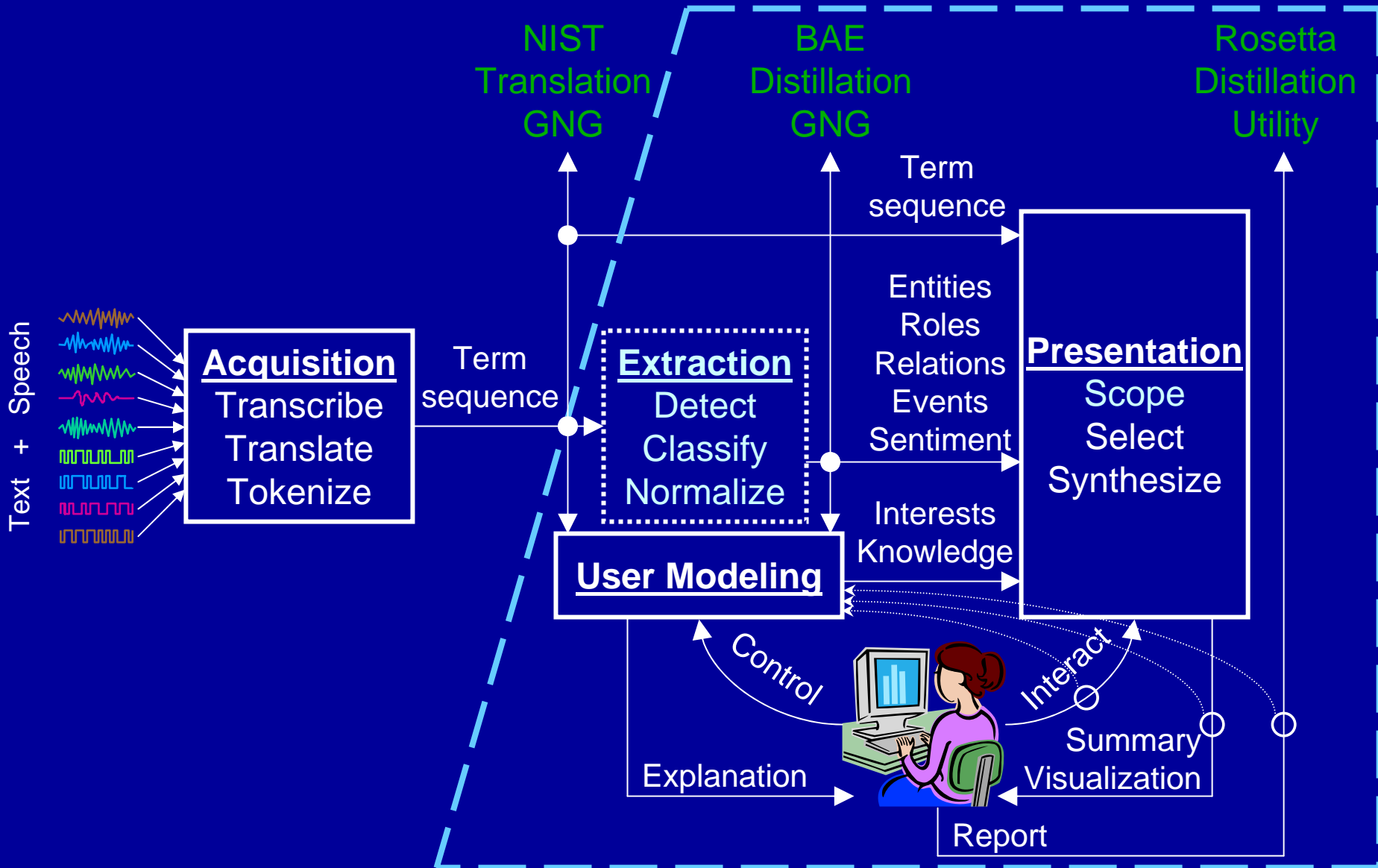
- **Mission:**

The goal of the GALE (Global Autonomous Language Exploitation) program is to develop and apply computer software technologies to **absorb, analyze and interpret huge volumes of speech and text in multiple languages, eliminating the need for linguists and analysts** and automatically providing relevant, distilled actionable information to **military command** and personnel in a timely fashion. Automatic processing "engines" will convert and distill the data, delivering pertinent, consolidated information in easy-to-understand forms to military personnel and monolingual English-speaking analysts in response to direct or implicit requests.

Intelligence Workflow



Rosetta



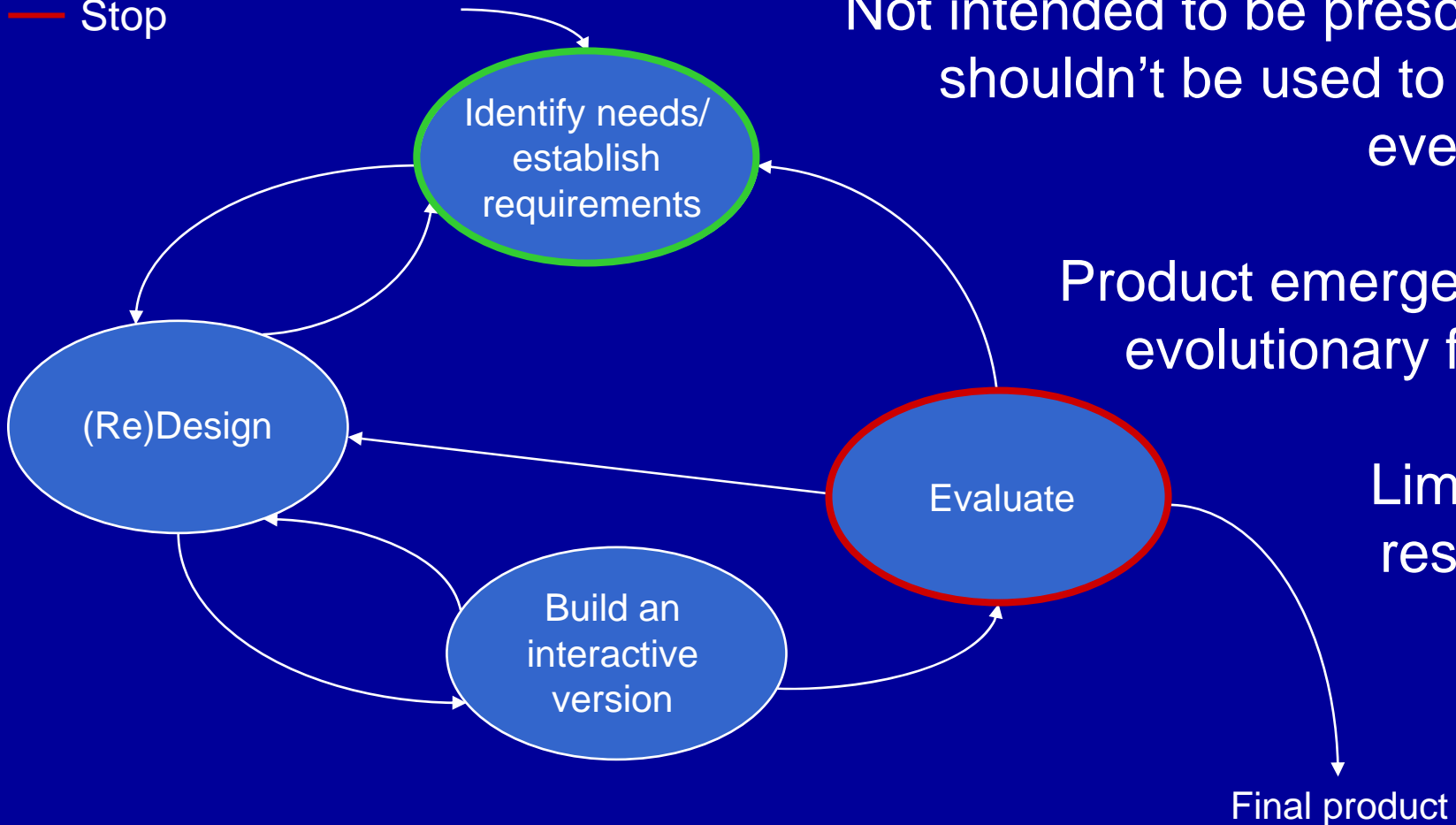
GALE Evaluations

1. Components (individual developers)
2. Go-no-go (BAE)
- 3. Formative intrinsic (CMU)**
- 4. Formative extrinsic (UMD)**
5. Utility (Aptima)
6. Insertion (real-world)

User-centric Development

— Start

— Stop

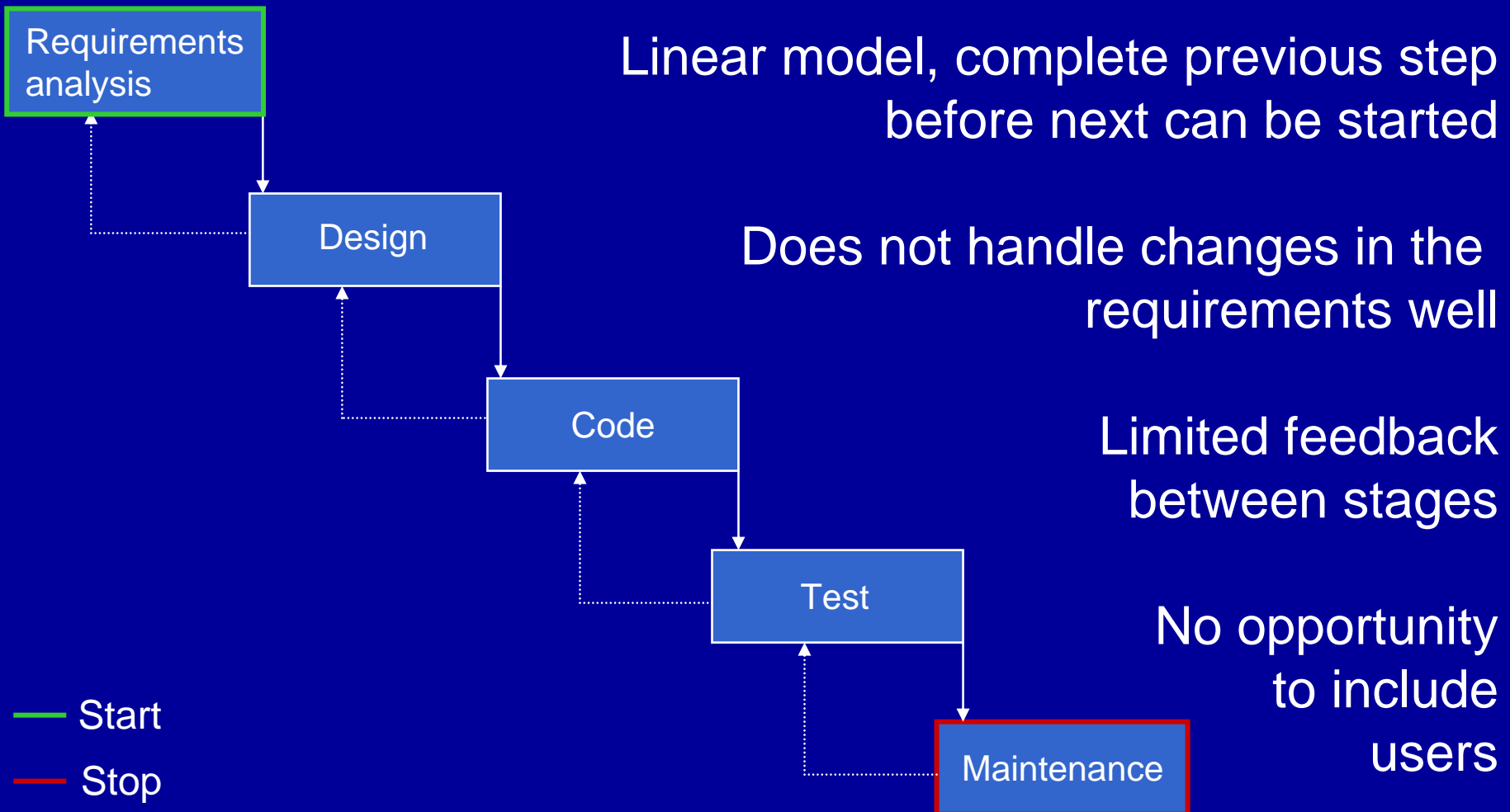


Not intended to be prescriptive,
shouldn't be used to design
everything

Product emerges in an
evolutionary fashion

Limited by
resources

Software engineering models: “Waterfall” lifecycle



Evaluation of IR Systems

- Traditional goal of IR is to retrieve *all* and *only* the relevant information in response to a single query
- All is measured by *recall*: the proportion of relevant IOs in the collection which are retrieved
- Only is measured by *precision*: the proportion of retrieved IOs which are relevant

Evaluation Problems

- Realistic IR is interactive; traditional IR methods and measures are based on non-interactive situations
- Evaluating interactive IR requires human subjects; the normal mode of evaluation is comparison between two systems (no gold standard or benchmarks); cannot compare a subject's searching on the same task in two systems
- Major tradeoffs between number of subjects and number of tasks; realism and control

Walk before you run

- 5 users, trained searchers (MLS students)
- Intelligence tasks (mostly)
- Live data: Arabic and Chinese TV news, news websites

Transcribe speech

Translate transcription

Index for retrieval

Web search-type interface



J2



Factual report

Login

- www.umdgale.com/gale
Username = *user#* (where # is 4-9)
Password = *rosetta*
- Rosetta
<http://rosetta.watson.ibm.com:9080/gale/>
- Use the advanced logged in
- Task = *jhutasks*
- Why have tasks?
 - Organize searching
 - Collaboration
 - Give the adaptive filtering half a chance

Options for viewing results

- Web results
 - Text view
 - HTML view
- Video Results
 - Play video
 - Storyboard
- Live Monitor

JHUTasks

Posted at www.umdgale.com/tinytasks

Rosetta URL again:

<http://rosetta.watson.ibm.com:9080/gale/>

Email Lynne LPLetten@umd.edu

Reaction Paper

- Improvements/ Feature requests
- Bugs
- Thoughts about the task
- Thoughts about GALE
- List of features ranked by importance
- Whatever

QUIZ

- Take the survey:
- <http://www.umdgale.com/galesurvey/index.php?sid=8>

GALE ASR & MT

- Earlier goals were for ASR and MT to allow English speakers to get the “gist” of non-English information
- GALE ASR and MT research has the goal of supporting broader use
 - It’s necessary to improve component performance,
 - But the broader goal is to support downstream automated processes (e.g. distillation) and to allow English speakers to understand incoming information quickly
- Focusing on new languages and genres
 - News, conversations, talk shows....
 - Arabic, Chinese, Korean, Farsi....

Challenges

- Moving from Scripted to Unscripted Speech
 - Untrained speakers (mumbling, misspeaking)
 - Dialect shifts (systems are tuned for a specific dialect)
 - Mixed language (interviewed person speaks in French with an interpreter simultaneously speaking in Arabic)
 - Incomplete sentences (many MT systems use sentence structure to help with the translation)
- Improving Machine Translation
 - How good does it have to be?
 - Any mistakes in ASR have implications here
- Distillation
 - Providing answers, not documents

How good is MT, practically?

How good does it need to be?

- Lots of numbers to measure MT performance, but it's hard for non-specialists to get much intuition from them.
- GALE is looking at new evaluation measures
 - Translation Error Rate – counts the edits required to make a machine translation convey the accurate meaning of the source document
 - Placing MT in the context of skill levels of human translators.
 - Usability experiments where humans use machine translated documents to complete tasks

Video Segmentation



2 minute segments in order to manage:

Size of file

Near real time access

Amount of text

2 minute segment split into 3 overlapping 60 sec segments



4 Kinds of Rosetta Duplication

some good, some bad

- Repetitive News Cycle (CNN Headline News)
- News “Teasers” (coming up next...)
- Multiple News Sources (CNN, ABC, AP, etc.)

X

- System Imposed (previous slide)

A Lot of Duplication!

Development Task

- 30 minutes
- Choose one or more:
 - 1. How can users develop search strategies to overcome ASR obstacles?
 - 2. How can users develop search strategies to overcome MT obstacles?
 - 3. How can duplication problems be avoided/minimized? Should they be?
 - 4. What might a better segmentation scheme look like?