Why combine these two areas in one lecture?
- MT evaluation often involves Human Factors considerations
- Human Factors is largely about evaluation
- “MT Evaluation is more studied than MT” -- Yorick Wilks, 1990
- E.g., LREC conference series!
- Because it is difficult to evaluate MT.

Why “manual”?
- “Manual” MT evaluation: versus newer automatic metrics (developed after Hutchins and Somers was written!)
- Automatic MT metrics will be covered next time.

But doesn’t BLEU etc. make all of this manual evaluation obsolete?
- Researchers need to validate/calibrate “core technology” metrics with human judgments
  - Human translation with same BLEU score is much better by human judgment
  - BLEU favors SMT over other MT approaches
- In real life, you need to know more than just “.434”
  - Does this particular MT system fit my particular needs?
- BLEU etc. are important: cheap/fast \(\rightarrow\) frequent!

Types of Evaluators
- Researchers
- Research sponsors
- Commercial developers, at several stages:
  - Prototype
  - Development
  - Operational
- Potential users (purchasers), of several types
  - Companies, government agencies
  - Assimilation vs. Dissemination
  - Translators
  - New: Casual web/PC users. But do they evaluate?
- Recipients of translations

Some Goals of MT Evaluation
- Commercial users’ goals:
  - Decide whether to use MT at all
  - Decide which MT system to buy
  - Track efficiency of MT vs. human translation
- Track quality of MT vs. human translation
- Researchers and Sponsors goals:
  - Measure quality of MT system as a whole
  - Analyze components of MT system
  - Track MT quality as system develops
Holy Grail of MT

- What MT users really want (pick any 2 of these 3):
  - Fully automated MT
  - General-purpose MT
  - Human-quality MT

-- Jaime Carbonell, early 1990s

What is “Quality of MT”?

- Or rather, what is quality of translation?
  - Semantic equivalence of source and target
  - Syntactic parallelism between S and T
  - Consistency (especially at lexical level)
  - Fluency of target text
  - Comprehensibility of target text for readers

What is “Quality of MT”?

Quality of MT

Or rather, what is quality of translation?

Semantic equivalence of source and target

Syntactic parallelism between S and T

Consistency (especially at lexical level)

Fluency of target text

Comprehensibility of target text for readers

Typical measures for translation

- “Accuracy”: Degree to which target preserves meaning (explicit and implicit) of source
- “Fidelity”: Degree to which target does not contain spurious content or nuance
- “Fluency”: Degree of well-formedness of target (trade-off with other measures!)
- “Literalness”: Degree to which target is literal translation vs. paraphrase
- Intelligibility, Style too

Recipient’s uses of translation

From Trujillo book:

- Scanning/“gisting”
- Content/“assimilation”
- Conversation/communication
- Publication/“dissemination”

Different uses require different levels of quality, and different system designs

Manual Evaluation Strategies

- “Black box” vs. “glass box”
- Test suite vs. test corpus

(Overall) output quality measurements

- US competitions: DARPA, TIDES, NIST, GALE
- Detailed linguistic error analysis (with classification)
- Task-based evaluation
- Total (“holistic”) system evaluation

US Gov. Quality Measures

- DARPA MT evals in 1990s:
  - Adequacy (is the meaning correct?)
    - Judged by bilingual humans, per chunk
  - Fluency (is output grammatical and fluent?)
    - Judged by monolingual humans, per sentence
  - “Informativeness”: subject answers SAT-test questions.
    - Now obsolete: strongly correlated with Adequacy
- GALE using edit-distance metric
Traditional Commercial MT: Evaluating MT for Dissemination

- Only human-quality translation acceptable.
- Human post-editing is therefore a “must”.
- Goal is to minimize total cost of translation: min(HT, MT+PE)
- Minimizing PE cost implies very different translation error model.

Evaluating MT for Dissemination

- Measure time gained, as efficiency factor
  \[ \text{Gain} = \frac{\text{Time}_{\text{Translate}}}{\text{Time}_{\text{Postedit}}} \]
- Challenge: Primadonna posteditors
  - Every translator wants to change MT output
    - They want to change other translator’s output too
  - There are many correct translations, all stylistic variants
  - Must train them to do minimum necessary

Minimizing PE cost

- Technical lexical/phrasal translation is key.
- Gross syntactic errors to be avoided (high cost to re-arrange text by editor).
- Inflection errors and other target-only errors are low cost (no reference to source)
- Glossing untranslated terms better than guessing

Comparative Evaluation of 4 Machine Translation Systems

<table>
<thead>
<tr>
<th>Translation Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English to Spanish</td>
</tr>
<tr>
<td>KANT</td>
</tr>
<tr>
<td>X</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Z</td>
</tr>
<tr>
<td>Heavy Postediting</td>
</tr>
<tr>
<td>Minimum Postediting</td>
</tr>
<tr>
<td>Fully Acceptable</td>
</tr>
<tr>
<td>Identical to Human</td>
</tr>
</tbody>
</table>

Traditional Commercial MT: “Holistic” system evaluation

- Software functionality, requirements, portability, reliability, etc.
- Extendibility/maintainability
- Personnel issues/Integration/Ease of use/Usability [Human Factors! See below]
- Cost/benefit analysis/efficiency [Business Case, later in this course]

CMU/LTI experience with manual MT evaluation

- KANT Catalyst system
- Pangloss (vs. IBM and Dragon)
- NESPOLE!
- Diplomat/Tongues
- Babylon/TransTac (MITRE)

- Mainly industry-funded or s2s MT systems
  - Difference between “core technology” and “total system” approach to research
Detailed evaluation example

- We will overview NESPOLE! evaluations as representative examples:
  - Evaluation of NESPOLE! Speech-to-Speech Translation system (human, sentence-level, quality-based, end-to-end + some components)
  - NESPOLE! task-based evaluation of speech-to-speech translation

NESPOLE! System Overview

- Human-to-human spoken language translation for e-commerce application (e.g. travel & tourism) (Lavie et al., 2002)
- English, German, Italian, and French
- Translation via interlingua
- Translation servers for each language exchange interlingua to perform translation
  - Speech recognition (Speech \(\rightarrow\) Text)
  - Analysis (Text \(\rightarrow\) Interlingua)
  - Generation (Interlingua \(\rightarrow\) Text)
  - Synthesis (Text \(\rightarrow\) Speech)

NESPOLE! User Interfaces

NESPOLE! Translation Monitor

Evaluation Types and Methods

- Individual evaluation of components
  - Speech recognizers
  - Analysis and generation engines
  - Synthesizers
  - IF (intercoder agreement, effectiveness)
- End-to-End translation quality
  - Sentence/segment level
  - From speech input to speech/text output
  - From transcribed input to text output
  - Architecture effectiveness: network effects
- Task-based evaluation
- User studies – what works and how well?
- Evaluating multi-modal interfaces

Single Component Evaluations

- Speech Recognizers:
  - Measure Word Error Rates (WERS) compared to a transcription of the input
- Analysis Modules: (from speech or text input)
  - Compare output from analyzer with manually annotated interlingua (IF) representations for the input
- Generation Modules: (from IFs)
  - Compare the generated output from IFs with the input utterance and assess quality of output
- Synthesizers: Does the output sound good?
### Example:

- G1: P VB K
- G2: P VB P
- G3: P VB K

#### Four point scale: “Perfect”/”OK”/”Bad”/”Very bad”
- OK: all meaning is translated, output somewhat disfluent
- Perfect: OK + output is fluent
- Bad: meaning is lost in translation
- Very Bad: completely not understandable
- Acceptable = Perfect + OK

#### End-to-End Evaluation Methodology

- 4 unseen dialogues for each language pair
- 2 from winter vacations, 2 from summer resorts
- Collected monolingually, 2 bilingually
- Monolingual and crosslingual evaluations
- Interlingua based evaluation at the Semantic Dialogue Unit (SDU)
  - Graded: one grader processed, all used segmentation
  - Calculate percent P/K/B/VB and Acceptable for each grader, average results across graders (or majority vote)

### Evaluation Results

<table>
<thead>
<tr>
<th>Language</th>
<th>M/R</th>
<th>SR Graded</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>63.3%</td>
<td>64.5%</td>
</tr>
<tr>
<td>German</td>
<td>63.5%</td>
<td>68.5%</td>
</tr>
<tr>
<td>French</td>
<td>71.2%</td>
<td>60.4%</td>
</tr>
<tr>
<td>Italian</td>
<td>76.5%</td>
<td>76.6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Transcribed</th>
<th>Speech Rec</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG-to-ITA</td>
<td>55%</td>
<td>45%</td>
</tr>
<tr>
<td>GER-to-ITA</td>
<td>32%</td>
<td>27%</td>
</tr>
<tr>
<td>ITA-to-ITA</td>
<td>46%</td>
<td>37%</td>
</tr>
<tr>
<td>ITA-to-GER</td>
<td>47%</td>
<td>31%</td>
</tr>
<tr>
<td>ITA-to-EN</td>
<td>40%</td>
<td>32%</td>
</tr>
<tr>
<td>ITA-to-FR</td>
<td>40%</td>
<td>32%</td>
</tr>
<tr>
<td>ENG-to-FR</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

### Task-based Evaluation: Motivation

- Accuracy-based evaluation can be very harsh (to get “K” ALL meaning must be preserved)
- Users studies indicate that with 65% accuracy we achieve almost 100% task completion
- But – this is highly dependent on the definition and complexity of the task!

### Task-based Evaluation of Speech-to-Speech Translation

- **Main issues:**
  - Focus on goal accomplishment rather than translation quality on a sentence/phrase level
  - Performing the evaluation requires human coding, based on a transcript of the conversation: inter-coder agreement
  - Score of a dialogue very sensitive to the flow of the dialogue, actions of the participants
  - Level of granularity of goals and sub-goals:
    - In this case, determined by the interlingua
Manual Evaluation Summary

- MT Evaluation is important for driving system development and the technology as a whole
- All aspects of a real system need to be evaluated – not just translation quality of individual sentences
- Human evaluations are costly, but are the most meaningful

Human Factors

11-731 Machine Translation
Bob Frederking
(Alex Rudnicky)

HF Introduction

- “Human factors” refers to how MT (and other) systems are designed so that actual human beings can use them.
- Also called “User-centered design”, CHI, HCI, MMI, HMI, OMI, UI Design, ergonomics, user experience
- Test the system early with real users executing real tasks!

“Usability Engineering”


HF Motivation

- HF analysis provides major cost savings for minor investments: a factor of 5 or 10!
- Plus, users like your programs better
- Save development effort: users may not want some features!
- Doesn't have to be a gigantic, expensive undertaking: 5-10% of budget, two person-years

Four biggest reasons for cost overruns:

- Frequent user change requests
- Overlooked tasks
- User's lack of understanding of own requirements
- Insufficient user/analyst communication
“Usability Engineering” slogans:
- Your best guess is not good enough.
- The user is always right: errors are not the user’s fault, they are your fault.
- The user is not always right: no vision, divergence between users.
- Users are not designers: too much customizing is bad.
- Designers are not users: you know too much.
- Vice presidents (or professors) are not users.
- Less is more.
- Details matter.
- Help doesn’t.

User Testing
- "The best is the enemy of the good" - Voltaire
- You don’t need elaborate laboratory, psychologists, videotapes, etc.
- User and task observation
- "Scenarios" (intersection of horizontal and vertical prototypes)
  - Cheap! Do them a lot! For every version.
  - Simplified thinking aloud: why user is doing it

“Heuristic Evaluation” checklist
- Simple and natural dialogue: relevant, frequent, logical order
- Speak the users’ language
- Minimize users’ memory load
- Consistency
- Feedback: people have a 10 second attention span
- Clearly marked exits
- Supply Shortcuts
- Good error messages: plain language, what’s wrong, what to do
- Prevent errors
- Help and documentation: task focused, well written

Apple online guide to HF
- Metaphors
- Reflect the User’s Mental Model
- Explicit and Implied Actions
- Direct Manipulation
- User Control
- Feedback and Communication
- Consistency
- WYSIWYG (What You See Is What You Get)
- Forgiveness
- Perceived Stability
- Aesthetic Integrity
- Modelessness
- Managing Complexity in Your Software

HF example: Pangloss
- Problem: Never really had real users!
- Fixed panes versus movable X windows.
- HAMT interface design decisions: Adjudicator vs. CMAT.
- CMAT user study resulted in MAT Tool paper.
- Benefit of HAMT proportional to skill of translator

HF example: Diplomat
- New branch on translation taxonomy: Communication. Low quality okay, if cheap, and if there is no other choice.
- Development of asymmetric interface, touchscreen.
- Carried out simple user test, with mocked-up, all-English “Wizard of Oz” interface.
- Try to visualize actual use of system:
  - Headset seemed like a bad idea
  - Handheld mike won’t work
  - Telephone handset!
- Similarly, Wearable needs a tablet.
- Importance of talking to real users: Croatian → Haitian!
  - Most Haitians cannot read, even Creole.
Summary

- For both HF design and MT evaluation, you must consider much more than the core technology.
- Golden rule of HF: Test early on real users really doing their real tasks!