Human-Computer Interaction

Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers.

Human-Computer Interaction

Human-Computer Interaction

Human-Computer Interaction

Computer Mouse

Human-Computer Interaction

Brief History of HCI

- HCI claims Alan Newell as the founding figure among others
- Alan Newell and Herb Simon were also pioneers of AI; first AI program called Logic Theorist to solve math theorems
- Turing award in 1975 for contributions to AI and human cognition
Human-Computer Interaction

Brief History of HCI

• Known in AI for work on natural language understanding; SHRDLU. Winograd Schema.
• Founded Stanford HCI group
• Advisor to Larry Page, Sergey Brin

Terry Winograd
NLP and Humans

Why should we care?
NLP and Humans

• We’ve mostly talked about NLP in isolation
• But at the end of the day NLP is about engineering tools
  • to be used by humans
    • for achieving their tasks
• This lecture is about:
  • Covering topics when humans and NLP models interact
  • Specifically, highlighting Issues that arise during interaction
NLP and Humans

- Evaluation
- Interaction
- Data Curation
- Social Biases
- Downstream Impacts
- Comp. Social Science
NLP and Humans

- Evaluation
- Interaction
- Data Curation
- Social Biases
- Downstream Impacts
- Comp. Social Science
Evaluation

• How to evaluate Natural Language Generation systems?

• Machine Translation
  • What makes a good translation?
    • The translation is grammatical and fluent? (Fluency, Grammaticality)
    • The translation preserves the meaning? (Adequacy)
    • The translation sounds natural? (Naturalness)
    • The translation uses local phrases and idioms? (Contextualness)
Evaluation

• How to evaluate Natural Language Generation systems?

• Machine Translation (Fluency, Adequacy)
Evaluation

• How to evaluate Natural Language Generation systems?

• Machine Translation (Fluency, Adequacy)

• Summarization (Meaning Preservation, Coherence)
Evaluation

- How to evaluate Natural Language Generation systems?

- Machine Translation (Fluency, Adequacy)
- Summarization (Meaning Preservation, Coherence)
- Story Generation (Relevance, Naturalness)
Evaluation

• How to evaluate Natural Language Generation systems?

• Machine Translation (Fluency, Adequacy)
• Summarization (Meaning Preservation, Coherence)
• Story Generation (Relevance, Naturalness)
• Dialog Agents (Appropriateness, Answerability)
## Evaluation

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Total</th>
<th>Criterion</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>40 (27%)</td>
<td>Readability</td>
<td>9 (6%)</td>
</tr>
<tr>
<td>Overall quality</td>
<td>29 (20%)</td>
<td>Appropriateness</td>
<td>7 (5%)</td>
</tr>
<tr>
<td>Informativeness</td>
<td>15 (10%)</td>
<td>Meaning preservation</td>
<td>6 (4%)</td>
</tr>
<tr>
<td>Relevance</td>
<td>15 (10%)</td>
<td>Clarity</td>
<td>5 (3%)</td>
</tr>
<tr>
<td>Grammaticality</td>
<td>14 (10%)</td>
<td>Non-redundancy</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Naturalness</td>
<td>12 (8%)</td>
<td>Sentiment</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Coherence</td>
<td>10 (7%)</td>
<td>Consistency</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>10 (7%)</td>
<td>Answerability</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Correctness</td>
<td>9 (6%)</td>
<td>Other criteria</td>
<td>124 (48%)*</td>
</tr>
</tbody>
</table>

Evaluation

• How to evaluate Natural Language Generation systems?

• In a previous lecture on Machine Translation:

**BLEU Score**

N-gram overlap between machine translation output and reference translation
Evaluation

• How to evaluate Natural Language Generation systems?

• In a previous lecture on Machine Translation:

  BLEU Score

  N-gram overlap between machine translation output and reference translation

  What does it capture?
Evaluation

- n-gram precision -> BLEU
- n-gram w/ synonym match -> METEOR
- tf-idf weighted n-gram -> CIDER
- n-gram recall -> ROUGE
- % of insert, delete, replace -> WER

- EDIT-DISTANCE

Evaluation

- n-gram precision -> BLEU
- n-gram w/ synonym match -> METEOR
- tf-idf weighted n-gram -> CIDER
- n-gram recall -> ROUGE
- % of insert, delete, replace -> WER

- EDIT-DISTANCE -> distance-based

Evaluation

- NLG evaluation can be done in several ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
Evaluation

• NLG evaluation can be done in 3 ways:

  • Untrained Automatic Metrics
    • BLEU, CIDER, METEOR, ROUGE
  • Machine Learning based Metrics
    • Sentence-Similarity, BERT-Score, BLEURT
Evaluation

- NLG evaluation can be done in 3 ways:
  
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT

Any flaws?
Evaluation

- NLG evaluation can be done in 3 ways:
  - Untrained Automatic Metrics
    - BLEU, CIDER, METEOR, ROUGE
  - Machine Learning based Metrics
    - Sentence-Similarity, BERT-Score, BLEURT
  - Human-centric Evaluation

Evaluation

• NLG evaluation can be done in 3 ways:

  • Untrained Automatic Metrics
    • BLEU, CIDER, METEOR, ROUGE
  • Machine Learning based Metrics
    • Sentence-Similarity, BERT-Score, BLEURT
  • Human-centric Evaluation  Most Preferred
Evaluation

• NLG evaluation can be done in 3 ways:
  • Untrained Automatic Metrics
    • BLEU, CIDER, METEOR, ROUGE
  • Machine Learning based Metrics
    • Sentence-Similarity, BERT-Score, BLEURT
  • Human-centric Evaluation

Evaluation

• NLG evaluation can be done in 3 ways:
  
  • Untrained Automatic Metrics
    • BLEU, CIDER, METEOR, ROUGE
  
  • Machine Learning based Metrics
    • Sentence-Similarity, BERT-Score, BLEURT
  
  • Human-centric Evaluation

Human Evaluation

- Human ratings are considered gold-standard in NLG evaluation
- Given a generated text how does a human rate it?

Human Evaluation

• Let’s try a sample human evaluation

On a scale of 1-5, rate the naturalness of the sentence

“Time flies like an arrow; fruit flies like a banana”

Human Evaluation

• Let’s try a sample human evaluation

How easy or difficult is the following sentence?

“Katie sipped on her cappuccino”

Human Evaluation

- Rating Scale popularly known as Likert scale
- Evaluation is outcome-level absolute assessment (OAA)

What are some issues with this approach?
Human Evaluation

- Rating Scale popularly known as Likert scale
- Evaluation is **outcome-level absolute assessment (OAA)**

- What are some issues with this approach?
  - **Interpretation**: What is meant by ‘naturalness’ or ‘difficulty’? How do you instruct annotators?
  - **Upper Bounds**: What does 1 and 5 mean?
  - **Interval width**: Is a jump from 3-4 same as 4-5?

Human Evaluation

• Another form: Comparative Ratings

Express preference for one of the following sentences S1 or S2

S1: “Hello world, I am Alexa”
S2: “Hey there, I am Alexa”

Human Evaluation

• Ranking system

• Evaluation is outcome-level relative assessment (ORA)

• What are some issues with this approach?
Human Evaluation

- Ranking system
- Evaluation is **outcome-level relative assessment (ORA)**

- What are some issues with this approach?
  - **Absolute Numbers**: What is the absolute performance of the model?
  - **Head-to-head Comparisons**: Massive number of comparisons

Human Evaluation
Issue: Language Subjectivity

- People find a particular tone to be better than the other
  - “Hey there” vs. “Hello World”

- What is toxic?
  - Depends on the person and their demographic group

- Given much of research happens (and data is collected) in West, these annotations can make NLP systems unworkable
  Ghosh, Sayan, et al. "Detecting cross-geographic biases in toxicity modeling on social media.” WNUT 2021
• For a long time, human ratings were gold standard

• However, with recent advances, humans find it difficult to distinguish between human-generated and model-generated text

## Human Evaluation

### Issue: Human Ratings

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall Acc.</th>
<th>Domain</th>
<th>Acc.</th>
<th>$F_1$</th>
<th>Prec.</th>
<th>Recall</th>
<th>Kripp. $\alpha$</th>
<th>% human</th>
<th>% confident</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stories</td>
<td>0.62</td>
<td>0.60</td>
<td>0.64</td>
<td>0.56</td>
<td>0.10</td>
<td>55.23</td>
<td>52.00</td>
</tr>
<tr>
<td>GPT2</td>
<td>*0.58</td>
<td>News</td>
<td>0.57</td>
<td>0.52</td>
<td>0.60</td>
<td>0.47</td>
<td>0.09</td>
<td>60.46</td>
<td>51.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.55</td>
<td>0.48</td>
<td>0.59</td>
<td>0.40</td>
<td>0.03</td>
<td>65.08</td>
<td>50.31</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>Stories</td>
<td>0.48</td>
<td>0.40</td>
<td>0.47</td>
<td>0.36</td>
<td>0.03</td>
<td>62.15</td>
<td>47.69</td>
</tr>
<tr>
<td>GPT3</td>
<td></td>
<td>News</td>
<td>0.51</td>
<td>0.44</td>
<td>0.54</td>
<td>0.37</td>
<td>0.05</td>
<td>65.54</td>
<td>52.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recipes</td>
<td>0.50</td>
<td>0.41</td>
<td>0.50</td>
<td>0.34</td>
<td>0.00</td>
<td>66.15</td>
<td>50.62</td>
</tr>
</tbody>
</table>

Table 1: §2 results, broken down by domain and model, along with the $F_1$, precision, and recall at identifying machine-generated text, Krippendorff’s $\alpha$, % human-written guesses, and % confident guesses (i.e., Definitely machine- or human-authored). * indicates the accuracies significantly better than random (two-sided t-test, for Bonferroni-corrected $p < 0.00333$).
Human Evaluation

• Types of human-involved evaluation
  • Intrinsic Evaluation (OAA, ORA)
    • Measure the text in itself
Human Evaluation

- Types of human-involved evaluation
  - Intrinsic Evaluation (OAA, ORA)
    - Measure the text in itself
  - Extrinsic Evaluation
    - Measure whether the system is able to help humans achieve a task

Human Evaluation

• Extrinsic Evaluation
  • Summarization -> Did the user get an idea of what a document was talking about?
  • Dialog Agents -> Was the user able to efficiently navigate through a website based on the outputs of a dialog agent?
  • Machine Translation -> Did the translation help user to achieve a task e.g., understanding directions and navigating in a foreign country?

Human Evaluation

• Extrinsic Evaluation
  • How?

  Evaluate at the system level and comparing systems that differ only in the NLG module

Human Evaluation

• Extrinsic Evaluation
  • Evaluate at the system level and comparing systems that differ only in the NLG module
  • Examples -
    • Reiter et al. (2003) generate personalized smoking cessation letters and report how many recipients actually gave up smoking.
    • Post-editing (Denkowski et al., 2014) can be used to measure a system’s success by measuring how many changes a person makes to a machine-generated text.

Human Evaluation

• Extrinsic Evaluation
  • **Most important**, as at the end of the day, it matters whether the end-user systems are usable
  • However,
    • Difficult to operationalize in NLP research
      • Systems are expensive to build and difficult to evaluate
    • Difficult to make progress within text generation
      • Systems used in varied context; other confounders in evaluation of systems other than just generated text
Human Evaluation

• Extrinsic Evaluation
  • HCI Research has several work that takes it to people and test it
  • Liebling et al. "Unmet needs and opportunities for mobile translation AI." CHI 2020.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking with people</td>
<td>I need to speak to someone who speaks another language than I do.</td>
</tr>
<tr>
<td>Getting around</td>
<td>I need to ask for directions but I don’t speak the local language.</td>
</tr>
<tr>
<td>Purchases</td>
<td>I need to buy something but I don’t speak the local language.</td>
</tr>
<tr>
<td>Checking into a hotel</td>
<td>I need to check-in to my hotel but I don’t speak the local language.</td>
</tr>
<tr>
<td>Get food</td>
<td>I need to buy food but I don’t speak the local language.</td>
</tr>
<tr>
<td>Dietary restrictions</td>
<td>I have a food allergy or preference I need to tell someone about but I don’t speak the local language.</td>
</tr>
<tr>
<td>Assistance</td>
<td>I need medical assistance but I don’t speak the local language.</td>
</tr>
<tr>
<td>Theft</td>
<td>I need help from the police but I don’t speak the local language.</td>
</tr>
<tr>
<td>Interpreter</td>
<td>I need a language interpreter or guide to help me communicate.</td>
</tr>
</tbody>
</table>

Table 1. Scenarios rated by respondents on dimensions of importance and frequency.
Interaction

- Human-Teacher, Machine-Learner
  - Learning from human feedback

- Machine-leading
  - Machines initiate interactions with their optimal competence, then humans respond with suggestions

- Human-leading
  - Humans initiate the task, then machines give suggestions based on their expertise

- Human-machine collaborators
  - Either can initiate. No explicit benefit for humans or machines

Interaction
Learning from human feedback

• Users generate rich signals that reveal model incorrectness and point to future model improvements (Krishna et al., PNAS 2022)

• Clickstream / Post-editing may implicitly reflect their expectations on a model like when they revise a model-generated text after accepting the suggestions

• How to integrate human feedback to improve the model itself?
• Also, called Human-in-the-loop (HITL)

Interaction
Learning from human feedback

Interaction
Learning from human feedback

Interaction
Learning from human feedback

• Making LMs bigger does not inherently make them better at following a user's intent.
  • untruthful, toxic, or simply not helpful to the user?
• InstructGPT
  • Fine-tune GPT-3 with labeler demonstrations of the desired model behavior (supervised learning)
  • Further fine-tune GPT-3 with dataset of rankings of model outputs (reinforcement learning with human feedback)

Interaction
Learning from human feedback

**Human Evaluation**

Evaluating Interaction

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Targets</th>
<th>Perspectives</th>
<th>Criteria</th>
<th>Social dialogue</th>
<th>Question answering</th>
<th>Crossword puzzles</th>
<th>Text summarization</th>
<th>Metaphor generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>First-person</td>
<td>Preference</td>
<td>Quality</td>
<td>Reuse</td>
<td>Ease</td>
<td>Enjoyment</td>
<td>Improvement</td>
<td>Enjoyment</td>
</tr>
<tr>
<td>Process</td>
<td>First-person</td>
<td>Preference</td>
<td>Quality</td>
<td></td>
<td>Helpfulness</td>
<td>Helpfulness Queries</td>
<td></td>
<td>Helpfulness</td>
</tr>
<tr>
<td>Process</td>
<td>Third-party</td>
<td>Preference</td>
<td>Quality</td>
<td></td>
<td>Queries</td>
<td></td>
<td>Edit distance</td>
<td>Queries</td>
</tr>
<tr>
<td>Output</td>
<td>First-person</td>
<td>Preference</td>
<td>Specificity</td>
<td>Interestingness</td>
<td>Fluency</td>
<td>Fluency</td>
<td>Consistency</td>
<td>Satisfaction</td>
</tr>
<tr>
<td>Output</td>
<td>First-person</td>
<td>Preference</td>
<td>Quality</td>
<td></td>
<td>Accuracy</td>
<td>Accuracy</td>
<td></td>
<td>Helpfulness</td>
</tr>
<tr>
<td>Output</td>
<td>Third-party</td>
<td>Preference</td>
<td>Quality</td>
<td></td>
<td></td>
<td></td>
<td>Consistency</td>
<td>Interestingness</td>
</tr>
</tbody>
</table>

Interaction
Collaboration and Design

• (Natural Language) Interfaces
• Communication of inputs / intermediate / outputs, their visualization
• Model Explanations
• Design choices:
  • name of the model (“GPT-3” vs. “Galactica”) (Khadpe et. al. CSCW)
  • preferences (what is an effective communication? politeness?)
Interaction
Bonus: Conceptual Metaphors


• Stereotype-content model: Warmth vs. Competence
  • Warmth: Follows assimilation theory
    • More warmth results in humans responding favorably
  • Competence: Follows contrast theory
    • More competence results makes humans not respond favorably
Thank you

Questions?

Icons from https://www.flaticon.com/