Putting the H into NLP

HCI + NLP Workshop
EACL 2021

Marti Hearst
UC Berkeley
April 22, 2021
What does HCI + NLP Mean?

- Using NLP to Help People within UIs
- Using HCI Techniques to Improve NLP
The User-Centered Design Process

3 Demos

Lessons on HCI + NLP for each
USER INTERFACE DESIGN IS AN ITERATIVE PROCESS

User Research
Design
Prototype
Evaluate
USER INTERFACE DESIGN IS AN ITERATIVE PROCESS

User Research
Design
Build
Evaluate
User Research

• Understand
  • Who users are
  • What their goals are
  • What tasks they need to perform.

• Task Analysis
  • Characterize what steps users need to take
  • Create scenarios of actual use
  • Decide which users and tasks to support.
Hi, what do you need?

- Requests & Payments
- Library
- Get Around
- Represent

Things To Do
- Public Works
- Vote

Make a Reservation
- Restaurant Name
- Name
- Date
- Time
- Phone
- Email
- Preferred Contact Form

Submit Request
**Why Prototype?**

- Experiment with alternative designs
- Fix problems before code is written
- Keep the design centered on the user
EVALUATION

• “Discount” techniques
  • Pilot tests on prototypes
  • Expert evaluations on prototypes

• Evaluations with participants
  • Task and interface specific
  • Designing these requires care, thought, and iteration
Typical HCI Process vs Typical NLP Process

HCI:
- Identify user need
- Develop method to address it
- Evaluate method on user needs

NLP:
- Identify NLP problem
- Develop algorithm
- Evaluate algorithm on accuracy, speed
Demo I:

ScholarΦ
Augmented Reader for Scientific Papers

CHI 2021 (to appear)
Motivation:
Have you ever struggled to keep track of notation or acronyms when reading a paper?

\[
\frac{1}{T} \sum_{t=1}^{T} \left[ \sum_{f=1}^{F} \log P(y_{ft}^{role} | P_G, V_G, \mathcal{X}) + \log P(y_t^{prp} | \mathcal{X}) + \lambda_1 \log P(\text{head}(t) | \mathcal{X}) + \lambda_2 \log P(y_t^{dep} | P_G, \mathcal{X}) \right]
\]

What does y stand for again?
Motivation:
Have you ever struggled to keep track of notation or acronyms when reading a paper?

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>He et al. (2018)</td>
<td>84.9</td>
<td>85.7</td>
<td>85.3</td>
</tr>
<tr>
<td>SA</td>
<td>85.78</td>
<td>84.74</td>
<td>85.26</td>
</tr>
<tr>
<td>LISA</td>
<td>86.07</td>
<td>84.64</td>
<td>85.35</td>
</tr>
<tr>
<td>+D&amp;M</td>
<td>85.83</td>
<td>84.51</td>
<td>85.17</td>
</tr>
<tr>
<td>+Gold</td>
<td>88.51</td>
<td>86.77</td>
<td>87.63</td>
</tr>
<tr>
<td></td>
<td>86.21</td>
<td>85.98</td>
<td>86.09</td>
</tr>
<tr>
<td></td>
<td>86.69</td>
<td>86.42</td>
<td>86.55</td>
</tr>
<tr>
<td></td>
<td>87.13</td>
<td>86.67</td>
<td>86.90</td>
</tr>
<tr>
<td></td>
<td>73.9</td>
<td>78.4</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>77.1</td>
<td>75.61</td>
<td>76.35</td>
</tr>
<tr>
<td></td>
<td>78.95</td>
<td>77.17</td>
<td>78.05</td>
</tr>
<tr>
<td></td>
<td>79.02</td>
<td>77.49</td>
<td>78.25</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 1: Precision, recall and F1 on the CoNLL-2005 development and test sets.
ScholarPhi Demo Video
Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols

Andrew Head • Kyle Lo • Dongyeop Kang • Raymond Fok
Sam Skjonsberg • Daniel S. Weld • Marti A. Hearst

Berkeley  Ai2  University of Washington  Alfred P. Sloan Foundation
Elaboration appears here
Usable Interfaces Require Multiple Rounds of Pilot Testing and Refinement

Four rounds of piloting
(In this case, the core idea was always positively responded to.)
Refined for months, continual feedback from relevant users
Studied the relevant HCI literature (reading, hypertext, etc.)
Real Applications Shed Light on Tough, Underexplored NLP Problems

Definition recognition is far from solved, and under-explored.

Complex, unsolved coordination problems.
Small Details Matter

Logical organization of information
Visible cues for current and next choices
Interactions that sustain flow
Legible text without artifacts
Standards followed
“In a live system, presentational details become disproportionately important. In our initial deployment, rendered text contained artifacts of the underlying tokenization …

These were no doubt relatively trivial matters of software engineering, but in initial informal evaluations, users kept mentioning these imperfections over and over again … distracting them from considering the underlying quality of the [system].”
Demo 2: WordZones
WORDZONES: IMPROVING WORD CLOUD DESIGNS
IEEE TVCG 2019

Marti Hearst, Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskoski, Steven Franconeri
language

words

systems

evaluation

NLP

rules

system

learning

algorithms

machine

input

text

NLP
data

tasks

transl

tion

intellige

tion
data

tasks

translation

tasks

examples

sentences

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

tasks

translation

ttasks
Word Clouds Are Engaging

They were originally intended for fun
However, today word clouds are frequently (mis)used in ...
What Is This About?

Is it a good way to show a summary of an upcoming course?
But they do not convey the intended message.

“Obama ... set out an agenda ... turning back the effects of climate change, launching a “moonshot” to cure cancer, and a grassroots movement to demand changes in the political system.”
HYPOTHESIS

Standard word clouds are detrimental to understanding.

A list of words is better for summarizing a text.
Hypothesis

Organizing the words both semantically and visually will improve understanding while retaining engagement.

How to prove this?
Can we fix this “informational” word cloud?
Let's Clean This Up
How to Test the Hypotheses?

Problems with Prior Evaluations

No shared datasets; no reproducibility

Often no human evaluation at all

Often unfair baselines, and only one baseline

Tasks often do not match underlying goals

“Find the largest word”
A NEW EVALUATION METHOD

**Goal:** Determine how well a layout “summarizes” the main topics of a document while retaining engagement

**Goal:** Reproducible evaluation

**Goal:** Evaluation the reflects the task
New Approach: based on the game of Taboo

• Build sets of words that unambiguously indicate a category
• The categories simulate the topics of a document
• See how long it takes someone to guess the underlying categories when the words are arranged in a cloud.
menu  waiter  dishes  tablecloth  bill

?  

restaurant

95% agreement or higher in isolation with native English speakers
These categories were time consuming to build; required rounds of interaction
moo  udder  professors
farm  students  dairy
milk  majors
dorms  campus
How many categories can you name in the time limit?
Can score from 0 to 5.
Hold some aspects constant, compare others
BENEFITS OF THIS EVALUATION APPROACH

- Reproducible
- Allows for fair comparisons across many baselines
- Yields statistically significant results
- Measures some of the underlying goals of the design
  - *(Pair it with subjective evaluations)*
Make sure the evaluation task matches realistic use cases.

Observe both subjective responses and quantitative results.
Do Compare Against a Strong Baseline

Compare against a full-featured baseline
Be sure the available contents are the same
Ideally vary as little between the two designs as possible
DO REPRODUCIBLE STUDIES

(This is a lesson for HCI from NLP)
Demo 3: News Chatbot
NewsLens
A Question Driven News Chatbot

Philippe Laban - phillab@berkeley.edu
John Canny and Marti Hearst
ACL 2020 System Track
**NewsLens Project**

- **Goals:**
  - Get people to engage in news
  - Help journalists

- **Approach:**
  - Analyze millions of news articles
  - Automatically detect stories, timelines, entities
  - Visualize these

- **Results:** Great NLP research in summarization, simplification, entity recognition, headline grouping. Creation of a highly valuable dataset for NLP

- **Problem:** Cool UI did not engage lay people or help journalists
Earlier Attempts at NewsLens Interface
Cambridge Analytica Scandal

Apr. 2018

Apr. 10, 2018: Facebook CEO Mark Zuckerberg is about to testify to Congress

Apr. 06, 2018: EU says Facebook confirmed data of 2.7 million Europeans 'improperly shared'

Apr. 04, 2018: Zuckerberg will testify before Congress on April 11

Mar. 27, 2018: Cambridge Analytica whistleblower: Vote Leave 'cheating' may have swayed Brexit referendum - video

Mar. 27, 2018: Facebook CEO Mark Zuckerberg will reportedly testify before Congress

Mar. 27, 2018: Zuckerberg will not appear before UK parliament committee

Mar. 26, 2018: FTC Confirms Facebook Probe Over Privacy Practices
Earlier Attempts at NewsLens Interface

Saudi crown prince

Actors

Mohammed bin Salman
King Abdullah
Mohammed bin Nayef
Barack Obama
Donald Trump
King Salman
Ibn Saud
Crown Prince
Prince Muqrin
Prince Sultan
Vladimir Putin
Prince Mohammed

Media

Saudi Arabia corruption arrests are a bold but risky attempt b...
Royal purge sends shockwaves through Saudi Arabia's elites
Senior Saudi royal ousted, princes reportedly arrested in pow...
Saudi Arabia Arrests 11 Princes, Including Billionaire Alwaleed ...
Saudi Arabia to begin issuing tourist visas
Saudi Arabia to allow women spectators in sports stadiums - ...
Saudi Arabia to allow women into sports stadiums starting in ...
Saudi Arabia Is Open for Business, but Not Everybody's Buyin...
The House of Saud is still in denial
Saudi Arabia Just Announced Plans to Build a Mega City That ...
Saudi Prince Tells Investors He's Taking on Religious Extremists
New Approach: NewsLens Chatbot

• **Goal**: Get casual news readers to engage more in-depth

• **Hypothesis**: Question-asking will increase engagement
NewsLens Chatbot Demo
News is organized into stories, each forming a chatroom.

Example stories: Iran Plane Crash, Boeing 737 MAX...
A chatroom starts with a timeline of events.

Each event is composed of multiple sources.
Questions are recommended to the reader.

Questions are updated as information is revealed.
Information is gathered from news and Wikipedia sources.
The reader can ask their own questions.

An extractive Q&A system finds a likely answer.
TOP STORIES

Iran Plane Crash
'Clerics get lost': Iran protests rage on for a third day
4 hrs ago

Iranian General Killing
Pompeo will not brief House committee on Iran, chairman says
4 hrs ago

British Royal Family
Queen agrees Prince Harry and Meghan can go it alone
5 hrs ago

US Democratic Election
Bernie Sanders told Elizabeth Warren in private 2018 meeting that a woman can't win, sources say
4 hrs ago

Trump Impeachment Inquiry
• **Goal**: Get casual news readers to engage more in-depth

• **Hypothesis**: Asking questions will increase engagement

• **Study Results**: When the chatbot recommends questions, news readers tend to have longer conversations

• **More Recent Work**: An audio podcast version finds that participants prefer a news podcast with Q&A vs one with stories read straight through
The unique NewsLens collection has enabled cutting-edge research in unsupervised summarization, simplification.
Develop something that looks cool.

However, it might not work as an interface for the intended users.
“After spending a lot of effort on the [cool NLP problem], the feedback we received from [our users] was quite sobering. Apparently, [our users] are a target audience with needs and preferences quite different from what computational linguists would prescribe.”

“To summarize, [the users] opted for a more intuitive presentation style…”
Summary: HCI + NLP

Using NLP to Help People within UIs

Using HCI Techniques to Improve NLP
How To Come Up With a Successful Novel User Interface Design?

• Be sure you’ve identified a real need
• Put user needs ahead of technology coolness
• Pilot test, pilot test, pilot test
• Small details matter as much as large ideas
How To Avoid Evaluation Errors?

• Don’t assess what makes one’s own technology look good versus assessing what people need or understand

• Don’t measure what is easy to measure versus what matters
**How To Devise a Good Evaluation?**

- Spend a lot of time thinking about it
  - *Read other experimental paper*
  - *Think deeply about the underlying goals of the application*
- Pilot test the measurement
  - *Refine it until it is getting consistent results*
THANK YOU!
Marti Hearst
Putting the H in NLP