CS329X: Human Centered NLP

Design Thinking

Diyi Yang
Stanford CS
Overview

**Motivation:** Why designs on top of NLP models are important

**Design thinking:**
- Double Diamond
- Problem reframing
- Prototyping
- Interviews and think aloud studies

Many slides credit to Sherry Wu
The biggest bottleneck for large language model startups is UX?
Motivating example: Build a Copywriting Tool

Build a tool that uses LLM to support copywriting

NLG task: users input the goal (e.g. advertising header) & provide context (e.g. product description), model outputs suggestions accordingly.
Motivating example: Build a Copywriting Tool

Build a tool that uses LLM to support copywriting

NLG task: users input the goal (e.g. advertising header) & provide context (e.g. product description), model outputs suggestions accordingly.

The biggest bottleneck for large language model startups is UX
https://www.copy.ai/
Motivating example: Build a Copywriting Tool

There are multiple tools that do this.

Especially in NLG & when we rely on existing LLM services, impossible/too time consuming to compare output quality!

More differentiable: the “periphery” of the product around the core language synthesis engine.

The biggest bottleneck for large language model startups is UX
https://www.copy.ai/
Motivating example: Build a Copywriting Tool

What aspects do you think would make a difference?

**How should a user provide the input context?**
How easy is it to understand how to provide this context, its purpose and impact on output?
Is it possible for the user to get really bad copy suggestions if they mess up the input context?

**How many suggestions should be shown to the user?**
Tradeoffs between variety vs. feeling overwhelmed?
How do you explain the idea of why there are multiple outputs?

**How do you rank suggestions to help users know which one they should pick?**
How do you explain what this ranking or score means to a user?
Should you “score” outputs at all?

**How do you allow users to balance their personal preferences with what the model thinks is optimal?**
Do you allow the user to “nudge” suggestions in a certain direction (tone, style, etc.)?
How do you avoid showing weird or odd suggestions which reduce user trust?

The biggest bottleneck for large language model startups is UX
<table>
<thead>
<tr>
<th>Question</th>
<th>Example for a LLM-based word editor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation</strong></td>
<td></td>
</tr>
<tr>
<td>What guarantees do I have about the system's accuracy?</td>
<td>Will the user have guarantees that anything the LLM suggests is grammatically sound? Could the LLM suggest fake words or provide false information? Could the system accidentally plagiarize a paragraph? To what extent is the user expected to “check” the output, and how does the user know when to check it?</td>
</tr>
<tr>
<td>How do I coach the system if it isn't doing what I want it to?</td>
<td>Should there be feedback mechanisms built into the system to allow the user to specify that they expected or wanted something different? What should a user do if the writing output is incongruous with their writing style, or if the system misinterpreted what they wanted (e.g. wrote a paragraph vs. finish a sentence)?</td>
</tr>
<tr>
<td><strong>Execution</strong></td>
<td></td>
</tr>
<tr>
<td>When should I ask the system for help?</td>
<td>Can the LLM be prompted at any time, or should the system indicate when it can be triggered? Are there situations where it is better vs. worse for the user to trigger the LLM (e.g. midway through a sentence, start of a paragraph)?</td>
</tr>
<tr>
<td>What should I ask the system?</td>
<td>Does the LLM purely function as an advanced form of auto-complete? Or, should the user be able to somehow specify or prompt the model to describe what they want written? Are there affordances or specific patterns for the user to indicate they want a paragraph filled out vs. their prior sentence rewritten vs. a tone change?</td>
</tr>
<tr>
<td>What can I expect the system to do for me?</td>
<td>What functionality should be built into the system - e.g. providing top level structure, finishing a sentence, rapid auto-complete for common phrases, style transfer, something else, fact checking? How do you make it clear to the user which of these things are possible vs. not?</td>
</tr>
</tbody>
</table>
Some Quick Take-aways

These issues really only surface once someone starts trying to use the product…

**This is how you go from “cool” to “useful.”**

These challenges are always present, regardless of system’s accuracy (within some bounds). Doesn’t matter if the LLM accuracy is 80% or 95%, the user still needs to reason through failure modes and understand what to expect when interacting with the system.
Overview of This Lecture

**Motivation**: Why designs on top of NLP models are important

**Design thinking**: 
User-Centered Design

“People ignore design that ignores people.” – Frank Chimero

User-centered design (UCD) is an iterative design process in which designers focus on the users and their needs in each phase of the design process.
Design Process: “Double Diamond”

“Double Dimond” is a typical design process. First Dimond finds the specific problem. Second Dimond finds the specific solution.

Divergent + convergent thinking:

Divergent: think broadly, keep an open mind, consider anything and everything
Convergent: think narrowly, bring back focus and identify 1-2 key problems / solutions.
Design Process: “Double Diamond”

**Discover:** Understand the issue rather than merely assuming it. It involves speaking to and spending time with people who are affected by the issues.

**Methods:**
- Multiple Perspective Framing
- Field studies
- Interviews
  
  ...
Design Process: “Double Diamond”

**Define:** The insight gathered from the discovery phase can help to define the challenge in a different way.

**Methods:**
- Task analysis
- Affinity diagrams.
...
“Within a design context, framing is often seen as the key creative step that allows an original solution to be produced. Designers report on the need to get to ‘the problem behind the problem’ (as initially presented by the client), and about creating a ‘fresh perspective.’”

– Bec Paton and Kees Dorst
Discover & Define: learn about users

These techniques focus on listening, observing and understanding the context in which people work and play.

They are exploratory and often open-ended, allowing for bottom-up analysis.

They include both small-scale qualitative techniques and quantitative data analysis.

<table>
<thead>
<tr>
<th>To...</th>
<th>Use...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understand users in their environment</td>
<td>Field studies: site visits, ethnography, or contextual inquiry to observe people doing their own tasks in their own setting.</td>
</tr>
<tr>
<td>Explore attitudes and expectations</td>
<td>Exploratory usability testing and interviews to collect information about their reactions to existing products or other conditions.</td>
</tr>
<tr>
<td>Know their goals and processes</td>
<td>Scenarios of use and other task analysis techniques to explore and document their workflow.</td>
</tr>
<tr>
<td>Identify quantitative demographics</td>
<td>Surveys on user demographics, product usage and other consumer habits.</td>
</tr>
<tr>
<td>Identify factors in the environment</td>
<td>Context of use audit to document environmental, social and access needs.</td>
</tr>
<tr>
<td>Create a portrait of users that captures what you have learned</td>
<td>Personas collect and document key aspects of different types of users.</td>
</tr>
</tbody>
</table>

Discover & Define: *learn the business environment*

These techniques focus on **what is happening in the business or personal domain**.

They are a snapshot of the competitive environment, trends surrounding the product and actual use of the product.

---


<table>
<thead>
<tr>
<th>To...</th>
<th>Use...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn about a new business environment</td>
<td><strong>Stakeholder interviews</strong> to collect input from different areas of the business domain.</td>
</tr>
<tr>
<td>Find trends or gaps in a business process</td>
<td><strong>Review problem reports</strong> from technical or customer support for usability problems or unmet needs</td>
</tr>
<tr>
<td>Understand usage patterns on a web site</td>
<td><strong>Traffic analysis</strong> of web site logs, looking for patterns in use, navigation, referrals and related sites or pages</td>
</tr>
<tr>
<td>Understand the competition</td>
<td><strong>Competitive audits</strong> or <strong>comparative usability test</strong> with competitive products, or other products and sites that are part of the business domain</td>
</tr>
</tbody>
</table>
Discover & Define: Interview

A method of asking questions & listening
Use planned interview protocol with open ended questions
Ask about what you can’t observe
Let people tell you what they know about themselves:
  What they do
  How they do things
  Their opinions on current activities
  How much they like one thing compared with another

“Go to the user, watch them do the activities you care about, and talk with them about what they’re doing right then.”

### Discover & Define: Interview

<table>
<thead>
<tr>
<th></th>
<th>Structured</th>
<th>Semi-structured</th>
<th>Unstructured</th>
<th>Focus group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-defined questions?</td>
<td>✓</td>
<td>✓</td>
<td>✘</td>
<td>✓</td>
</tr>
<tr>
<td>Open-ended questions?</td>
<td>✘</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Fixed order of questions?</td>
<td>✓</td>
<td>✘</td>
<td>✘</td>
<td>✘</td>
</tr>
<tr>
<td>Fixed number of questions?</td>
<td>✓</td>
<td>✘</td>
<td>✘</td>
<td>✘</td>
</tr>
<tr>
<td>Can ask additional questions?</td>
<td>✘</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Semi-structured is **most common**.

Allows for **exploratory** studies.

Provides comparable, reliable data, and the flexibility to ask follow-up questions.

https://www.scribbr.com/methodology/semi-structured-interview/
Semi-Structured Interviews: Thematic analysis

Identify common themes from transcriptions - topics, ideas and patterns of meaning that come up repeatedly.

Define codebook, multiple coders, compute annotator agreement.

<table>
<thead>
<tr>
<th>Interview extract</th>
<th>Codes</th>
</tr>
</thead>
</table>
| Personally, I'm not sure. I think the climate is changing, sure, but I don't know why or how. People say you should trust the experts, but who's to say they don't have their own reasons for pushing this narrative? I'm not saying they're wrong, I'm just saying there's reasons not to 100% trust them. The facts keep changing - it used to be called global warming. | • Uncertainty  
• Acknowledgement of climate change  
• Distrust of experts  
• Changing terminology |
How many participants to interview?

Depends on Goals, Context, Resources/Timing.
As many as you

need for finding new things out (data saturation)

can afford (time, incentives, etc.)

have time to analyze (2x+ per participant)

**Magic:** 12 is a good number (minimum of five)

Make sure to choose representative users

**Or stop when findings start to converge**

---

Discover & Define: Task & Information analysis

These techniques focus on the information or actions that users will need to meet their goals.

<table>
<thead>
<tr>
<th>To...</th>
<th>Use...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explore a group process or work flow</td>
<td>Participatory design, or PANDA (Participatory Analysis and Design Activities) techniques, such as The Bridge, to develop a consensus view of the overall process.</td>
</tr>
<tr>
<td>Learn about relationships between information or tasks</td>
<td>Card sorting to create logical groups from the users’ point of view</td>
</tr>
<tr>
<td>Decide how to organize a task or collection of information</td>
<td>Affinity diagrams, navigation flow charts to group and explore the structure of the information</td>
</tr>
</tbody>
</table>

Case: Improve Word Editor

Add intelligent language functionality into a Word document editor, to improve individual users’ writing experience.

Dear Educator,

[intro paragraph that isn't too cheesy]

Part-time Hebrew schools serve a vital role in the continuation of Jewish cultural literacy in America. Over [reference: 80%? PEW?] of self-identified Jews’ primary, if not only, exposure to Jewish education is in 20-25 weeks per year, not Jewish day schools. "First Century Skills" of students the future work, we have a responsibility to help prepare them for the workplace.

Jewish educational attainment around the world | Pew Research Center
http://www.pewforum.org/2016/12/13/jewish-educational-attainment/

Eight facts about Orthodox Jews from the Pew Research survey | Pew Research Center
www.pewresearch.org/.../eight-facts-about-orthodox-jews-from-the-pew-research-survey...

More Search on  Jewish  ×  Education  ×  Part-time  ×  

explore, debate, process and re-process ideas – why are we not jumping onto that bandwagon with glee?
Reframing

Figure 2: Left: A developed version of the Notebook (Figure1), used as boundary object between HCI and NLP researchers. “Contexts” that can help inform intelligent function outputs are marked blue. Middle and Right: Reframing the problem of designing writing assistance as other canonical NLP technical problems. This expands our design space to the intersection between what authors want and what existing NLP capabilities can do.
Use reframing to drive interviews (1/2)

Framing Writing Assistance as Conversational AI

**Reframing purpose – to find new design questions:**
Whom would authors like to talk to and for what purpose?
What information can conversational assistance offer?

**Findings:** Participants...
Pick those who are close to their target readers as their "beta-readers".
Write to meet the expectations and needs of their target readers.
Read documents from their target venue to infer the expected length, lexical complexity, or level of detail.

**Transfer to function: “ask your reader”**
Mines documents from an author-identified venue. The author can request insights about these documents or make comparisons between their own writing against it.

"Am I writing too formally?"
"How long is a typical introduction section in [venue]?"]
Use reframing to drive interviews (2/2)

Reframing purpose – for near-future technical possibility:
Search is a relatively matured NLP sub-domain. how do authors sought information during writing?

Findings: Participants…
Search for sample rhetorical structures for reference, e.g. “[quotation mark][comma] in comparison to [quotation mark]”

Does not work: Current search focus on content, not structure.
Read documents from their target venue to infer the expected length, lexical complexity, or level of detail.

Transfer to function: “rhetorical search function”
Find texts similar in language structure and composition to the query.
Design Process: “Double Diamond”

**Develop:** Give different answers to the clearly defined problem, seeking inspiration from elsewhere and co-designing with a range of different people.

**Methods:**
- Storytelling;
- Minimum Viable Product;
- Rapid prototyping....
Develop: Evaluate designs-in-progress, Via formative usability testing

**Formative usability testing**: test with representative users and representative tasks on a representative product.

To not only evaluate a product or prototype, but to provide recommendations to improve it.

<table>
<thead>
<tr>
<th>To...</th>
<th>Use...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explore different solutions</td>
<td>Paper prototyping, and task-based or persona walk-throughs to explore the structure of the information</td>
</tr>
<tr>
<td>Collect informal input</td>
<td>Informal testing and hallway reviews to collect rapid input for sections of a design</td>
</tr>
</tbody>
</table>

What’s a prototype?

Physical realizations of the research and design process in a tangible form.
Can be used to get a sense of what it would be like to experience the product/service.
Can appear at varying levels of fidelity
- Paper, low-fidelity prototypes usually show up at earlier stages in the process
- Higher-fidelity prototypes show up later
Common prototyping methods

**Wizard-of-Oz**
Fake features so that the user thinks that the responses are computer-driven when they are actually human-controlled.
*Challenge for NLP: AI errors are hard to simulate.*

**Mimic simple functionality**
Challenge for NLP: cannot simulate SOTA model capabilities

**Ensemble multiple simple models and expectations.**
(More recent) *use large language models.*

---

**Figure 3:** This prototype interface is a simple text editor. At any time of their writing, users type @ to signal the start of an intelligent function request and Enter to end. When they click on a request, intelligent assistance pops out. This prototype probes users’ needs and wants for writing assistance, and their reactions to the simulated intelligent responses.
Prototyping with AI/NLP: Persona

**Algorithmic persona**: human roles that users assign to the algorithm to explain the algorithm’s goals, behaviors, and characteristics.

**Example**: YouTube recommendation algorithm
65 years of video are uploaded every day...
The way YouTube content creators perceive these algorithms affect their attitudes and

**Recommendation Algorithm as Agent**

**Agent:** manages and helps creators in their work by finding an audience for them and promoting them.

“YouTube will favor you in the algorithm, which would then lead to more views and more subscribers.”

**Blessed by, Build a relationship with, To please, Work with**

“You wanna be friends with the YouTube algorithm which decides to push your video or not.”
Recommendation Algorithm as Gatekeeper

**Gatekeeper:** stands between content creators and the viewers and determines whether YouTubers’ content gets viewed.

“...there is [an] algorithm between you and the viewers. You need to try to understand the algorithm and play to its strengths, or kinda get really lucky.”

**Bribe, Circumvent, Fit in**

“I ended up getting a lot of views because I actually piggybacked a very popular trend at the time.”
Recommendation Algorithm as Drug Dealer

**Drug dealer:** keeps viewers *addicted* to the platform. “The algorithm is really good at keeping us here.”

**Rebel, Complicit, Addictive**

“My model is slow disperse growth. I’m trying to go the other way against the click-bait, viral algorithm. My goal is to not to follow that model. It is a dangerous path -- it’s luck.”
Prototyping with AI/NLP: Persona to behavior

The use of persona: Describe roles that are familiar, use them to guide design. What would be your expectations on a model if it’s introduced as agent, gatekeeper, etc.?
Prototyping with AI/NLP: Persona to behavior

**Agent:** Contracts between YouTubers and the algorithms?

**Gatekeeper:** Creators ask the algorithm to explain/appeal why their video got demonetized?

**Drug dealer:** address the public health concerns of algorithm as a drug dealer?
Mixed initiative systems allow users to interact with them in a collaborative way, where the user and the system both take an active role in carrying out tasks or making decisions.

Advocates elegant coupling of **automated services** with **direct manipulation**.

“Autonomous actions should be taken only when an agent believes that they will have greater expected value than inaction for the user.”

Principles for Mixed-initiative user interfaces

Developing significant **value-added** automation (v.s. direct manipulation)

Considering **uncertainty** about a user’s goals

Considering the status of user’s **attention** (minimize distraction, cost vs. benefit of deferring action)

Inferring **ideal action** in light of costs, benefits and uncertainties (expected values of actions!)

Employ **dialog** to resolve key uncertainties (interactions!)

Allowing efficient **direct** invocation and termination

Minimizing the **cost of poor guesses** about action and timing

**Scoping precision** of service to match uncertainty, variation in goals – do less if uncertain!

Providing mechanisms for efficient agent-user collaboration to **refine** results

Employing **socially appropriate** behaviors for agent-user interaction

Maintaining working **memory** of recent interactions

Continuing to **learn** by observing (e.g., about user’s goals, etc.)

Project Tips
Pick a question that you’re excited about

Broadly relevant to HCI + NLP

✦ Could you formulate a research question to deeply explore it?
✦ What type of data might be available for you to use?
✦ Which softwares or tools could you use to work on it?
✦ How do you evaluate the outcome of your project?
Form Research Group

Posted as an assignment on Canvas, dues on 4/18

You will fill in a short Google Form that documents your group members, and a general description of your project. Your group should contain 1-3 people.

In the form, you will answer these questions:

1. **The problem:** what are you trying to do
2. **Why bother:** Summarize why the problem is important, or why we care about solving it.
3. **Status-quo:** Current solutions and why they may fail
4. **Optional** [Your proposed method]: If I had a solution, what would it look like?
5. **Optional** [Evaluation / metrics of success]: How do I know if I solved the problem?

If you are looking for project partners, please post to Ed Discussion!
What’s a good project?

It should generally be relevant to HCI+NLP.

Just pick projects that interest you!
  TA and I will reach out if a project seems too far off

But, make sure to do a small scoped project that’s suitable for a quarter

And, if you want the course to count as your technical requirement, you might want to choose a project that require more coding practice.
Resources to check out

Top course projects sometimes end up into actual paper submissions to either full conferences or workshop venues.

Checking out workshop papers published in (*some of them focus on general AI):

- HCI+NLP @ NAACL 2022
- HCI+NLP @ EACL 2021
- Human Evaluation of Generative Models @ NeurIPS 2022
- In2Writing @ CHI 2023
- InterNLP @ NeurIPS 2022
What could be a final project?

Some sample topics could be:

**System / interface**: Designing and evaluating a natural language interface for a mobile or web application, with a focus on usability and user experience.

**Analysis**: Examining the biases present in a specific NLP model or dataset, and designing solutions to mitigate those biases.

**Design**: Analyzing and visualizing model decisions (e.g. interpretability) to accommodate the needs of specific domain experts.
Key Considerations

Availability of data

Be careful in deciding whether to collect and annotate your own data

ML framework

Huggingface, sklearn, keras, pytorch, Tensorflow

Statistical models

R, Stata, etc.

Availability of computation
Literature Review

Conduct a thorough literature survey

A few places to check out:
  Google Scholar
  ACL Anthology (https://aclanthology.org/)
Types of Projects

Visualization or interpretability analyses of neural networks
Apply/extend a computational NLP method to real world problem
Develop new methodologies to leverage human feedback/preferences
Fairness, bias, or ethical issues around existing NLP tools
Improve existing NLP pipelines
Building interactive NLP systems to allow humans to interact with LLMs
Simulating personas via LLMs
NLP for social good (e.g., accessibility, climate change, etc)
Position papers or a critic (talk to us first)
Recommendations for Successful Projects

Start early and work on it every week rather than rushing at the end

Get your data first!

Have a clear, well-defined research question (novel/creative ones ++)

Results should teach us something

Visualize results well

Divide the work between team members clearly
Common Issues

Data not available or hard to get access to

No code written for model/data processing

Team starts late

Results/Conclusion don’t say much besides that it didn’t work
   Even if results are negative or unexpected, analyze them
Resources

Computation
  Google Cloud/Google Colab

Discussion
  Come to TA and Diyi’s Office hours
Come up with your own idea and talk to us!
Principles for Mixed-initiative user interfaces

Developing significant value-added automation (v.s. direct manipulation)
Considering uncertainty about a user’s goals
Considering the status of user’s attention (minimize distraction, cost vs. benefit of deferring action)
Inferring ideal action in light of costs, benefits and uncertainties (expected values of actions!)
Employ dialog to resolve key uncertainties (interactions!)
Allowing efficient direct invocation and termination
Minimizing the cost of poor guesses about action and timing
Scoping precision of service to match uncertainty, variation in goals – do less if uncertain!
Providing mechanisms for efficient agent-user collaboration to refine results
Employing socially appropriate behaviors for agent-user interaction
Maintaining working memory of recent interactions
Continuing to learn by observing (e.g., about user’s goals, etc.)

Case Study: Interactive Machine Translation

Predictive Translation Memory

La Cour Suprême des États-Unis confirma en 2008 la constitutionnalité de la loi de l’Indiana.

The U.S. Supreme Court upheld in 2008 the constitutionality of the Indiana law.

Les autorités républicaines s’empressèrent d’étendre cette pratique à d’autres États.

The Republican authorities if empressèt extending this practice to other states.

Au cours des deux dernières années, elles parrainent des projets de loi dans 54 États pour forcer les électeurs à présenter une carte d’identité avec photo.

Over the past two years, they are sponsoring bills in 34 states to force voters to present a photo identification.

Il est important de noter que, contrairement au Québec, les citoyens américains ne disposent pas de carte d’identité universelle comme la carte de l’assurance maladie.

It is important to note that, unlike in québec, the American people do not have universal identity card as the health insurance card.

De fait, 11% des citoyens américains, soit 21 millions de personnes en âge de voter, ne possèdent pas de cartes d’identité avec photo émises par une agence gouvernementale de leur État.

Experiment will end in 2:46 without continued translations.
PTM recap: Rationals for seemingly simple decisions

**Design:** Re-use familiar hotkeys e.g., CTRL+Enter Typing activates interactions
   Translators are fast typists: want to avoid the mouse

**Design:** One column, interleaved layout
   Translators read (20-25% of translation session), 2-column will be cumbersome

**Design:** Text color encoding
   Ownership: AI can’t modify human text, human can accept but not modify AI text

**Principle:** Horvitz #6 - Employing socially appropriate behaviors for agent–user interaction.
PTM recap: Source comprehension

**Design:** highlight translated words

**Principle:** Horvitz #11 - maintaining working memory of recent interactions
PTM recap: Source comprehension

**Design:** highlight translated words

**Principle:** Horvitz #11 - maintaining working memory of recent interactions

**Design:** allow for word-to-word query

**Principle:** Horvitz #6 - allowing efficient direct invocation and termination
PTM recap: Target gisting

**Design:** Full best translation  
**Principle:** Horvitz #10 - Employing socially appropriate behaviors for agent-user interaction

**Design:** Real-time updating  
**Principle:** Horvitz #9 - providing mechanisms for efficient agent-user collaboration to refine results termination

- Les étudiants eux-mêmes n'ont pas les moyens de se rendre à des cours, nous essayons de les aider de cette manière.
- The students themselves cannot be required to attend courses, we are trying to help them.
- Dans le cadre de l'institut Jedlička, nous transférerons ce projet dans un autre.
PTM recap: Target generation

**Design:** Insert complete translation

**Principle:** Horvitz #6 - allowing efficient direct invocation and termination
PTM recap: Target generation

**Design:** Insert complete translation

**Principle:** Horvitz #6 - allowing efficient direct invocation and termination

**Design:** Real-time autocomplete dropdown

**Principle:** Horvitz #5 - employing dialog to resolve key uncertainties
PTM recap: Other principles?

- Developing significant value-added automation (v.s. direct manipulation)
- Considering uncertainty about a user’s goals
- Considering the status of user’s attention (minimize distraction, cost vs. benefit of deferring action)
- Inferring ideal action in light of costs, benefits and uncertainties (expected values of actions!)

Employ **dialog** to resolve key uncertainties (interactions!)

Allowing efficient **direct** invocation and termination

Minimizing the cost of poor guesses about action and timing

Scoping precision of service to match uncertainty, variation in goals — do less if uncertain!

Providing mechanisms for efficient agent-user collaboration to **refine** results

Employing **socially appropriate** behaviors for agent-user interaction

Maintaining working **memory** of recent interactions

Continuing to learn by observing (e.g., about user's goals, etc.)
Principles for Mixed-initiative user interfaces

- Developing significant **value-added automation** (v.s. direct manipulation)
- Considering **uncertainty** about a user’s goals
- Considering the status of user’s **attention** (minimize distraction, cost vs. benefit of deferring action)
- Inferring **ideal action** in light of costs, benefits and uncertainties (expected values of actions!)
- Employ dialog to resolve key uncertainties (interactions!)
- Allowing efficient direct invocation and termination
- Minimizing the **cost of poor guesses** about action and timing
- **Scoping precision** of service to match uncertainty, variation in goals – do less if uncertain!
- Providing mechanisms for efficient agent-user collaboration to refine results
- Employing socially appropriate behaviors for agent-user interaction
- Maintaining working memory of recent interactions
- Continuing to **learn** by observing (e.g., about user’s goals, etc.)

PTM recap: Other principles?

- Developing significant value-added automation (v.s. direct manipulation)
- Considering uncertainty about a user’s goals
- Considering the status of user’s attention (minimize distraction, cost vs. benefit of deferring action)
- Inferring ideal action in light of costs, benefits and uncertainties (expected values of actions!)
- Employ dialog to resolve key uncertainties (interactions!)
- Allowing efficient direct invocation and termination
- Minimizing the cost of poor guesses about action and timing
- Scoping precision of service to match uncertainty, variation in goals – do less if uncertain!
- Providing mechanisms for efficient agent-user collaboration to refine results
- Employing socially appropriate behaviors for agent-user interaction
- Maintaining working memory of recent interactions

**Continuing to learn by observing (e.g., about user’s goals, etc.)**

**RLHF :)**
Design Process: “Double Diamond”

**Deliver:** Involves testing out different solutions at small-scale, rejecting those that will not work and improving the ones that will.

**Methods:**
- Survey;
- Think Aloud;
- Usability testing;
  ...

- Discover
  - Research
  - General Problem
- Define
  - Synthesis
  - Specific Problem
- Develop
  - Ideation
- Deliver
  - Implement
  - Specific Solution
Develop & Deliver: Evaluate usability results

These techniques include both measuring the success of a design (against usability performance and satisfaction criteria) and establishing benchmarks metrics. They require a more formal test protocol, and realistic tasks.

<table>
<thead>
<tr>
<th>To...</th>
<th>Use...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine whether a product is meeting its usability goals</td>
<td><strong>Summative usability testing</strong>, measuring performance against criteria (and possibly benchmark values)</td>
</tr>
<tr>
<td>Learn how a product compares to its competitors</td>
<td><strong>Comparative testing</strong> with the same tasks performed using two or more products</td>
</tr>
<tr>
<td>Find out whether users like a product</td>
<td><strong>Satisfaction surveys</strong>, as part of a usability test or with random users.</td>
</tr>
<tr>
<td>Test a design against scenarios of use</td>
<td><strong>Usability testing</strong> in formal or informal settings:</td>
</tr>
<tr>
<td>Understanding what parts of the interface draw the user’s visual attention</td>
<td><strong>Eyetracking</strong> lets you see exactly where a user looks on the screen, and for how long.</td>
</tr>
<tr>
<td>Ensuring access for all</td>
<td><strong>Usability testing</strong> with people with disabilities</td>
</tr>
</tbody>
</table>

What is a “Think Aloud?”

A research method used to gain insight into a person's thought processes as they perform a task or solve a problem. The participant is asked to verbalize their thoughts as they perform the task, which allows the researcher to understand how the participant approaches the task. "Thinking aloud may be the single most valuable usability engineering method."

“I’m going to ask you to ____ and while you are doing that, can you tell me whatever you are thinking. Whatever comes into your mind while you are working on that. Okay?”

**Protocol**

- Give participants specific tasks to accomplish (but not HOW to do it)
- Have them speak aloud as they complete the tasks
- Keep interruptions to a minimum
- Ask for open-ended questions & clarification after the task is complete
- Learning effect - if you make tasks, watch for biasing test due to order
- Typically used to test the usability of a website, app or object

What is a “Think Aloud?”

Think-aloud user studies are a mixer of **quantitative** and **qualitative** studies.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Quantitative</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>Gather numerical data to be analyzed using statistical methods</td>
<td>Gathering descriptive, non-numerical data to be analyzed through interpretation and contextualization</td>
</tr>
<tr>
<td><strong>Data source</strong></td>
<td>surveys, questionnaires, experiments</td>
<td>interviews, observations, and document analysis</td>
</tr>
<tr>
<td><strong>Presentation</strong></td>
<td>tables, graphs, and statistics</td>
<td>quotes and narratives that reflect the participants' experiences and perspectives</td>
</tr>
<tr>
<td><strong>Goal</strong></td>
<td>establish cause-and-effect relationships between variables</td>
<td>gain a deeper understanding of social phenomena, meanings, and processes</td>
</tr>
</tbody>
</table>
Advantages of think aloud studies

**Rapid, high-quality, qualitative** user feedback

Data available from **range of sources:**
- Direct observation of what the subject is doing.
- Hearing what the subject wants, or is trying, to do.

If participant gets into difficulties, observer has the chance to **clarify situation**

High flexibility; experiment may easily be steered by the observer

In person allows meaningful, direct dialogue
We present Predictive Translation Memory, an interactive, mixed-initiative system for human language translation. Translators build translations incrementally by considering machine suggestions that update according to the user’s current partial translation.

PTM: Experimental Design

Comparative analysis
“We compared our system to post-editing, which is a strong baseline [29, 21], and is also the most common commercial use of MT.”

Clear research questions
Time - PTM faster than post-edit?
Quality - PTM == better translation?

Task   translate French→English or English→German
Source Text  ≈3,000 tokens of News/Medical/Software
Conditions  post-edit (pe) and PTM
Expert Subjects  16 per language pair
RQ1: Time – PTM faster than post-edit?

Metric: log of time (more tolerant of outliers)
Quantitative analysis (find robust evidence)
  Compare mean (for general understanding)
  Linear mixed effects models (for understanding significance, important factors)

<table>
<thead>
<tr>
<th></th>
<th>Fr-En</th>
<th></th>
<th>En-De</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sign</td>
<td>p</td>
<td>sign</td>
<td>p</td>
</tr>
<tr>
<td>ui (PTM)</td>
<td>+</td>
<td>o</td>
<td>+</td>
<td>★★</td>
</tr>
<tr>
<td>ui order</td>
<td>−</td>
<td>★</td>
<td>−</td>
<td>★★</td>
</tr>
<tr>
<td>normalized edit distance</td>
<td>+</td>
<td>★★★</td>
<td>+</td>
<td>★★★</td>
</tr>
<tr>
<td>no edit (True)</td>
<td>−</td>
<td>★★★</td>
<td>−</td>
<td>★★★</td>
</tr>
<tr>
<td>gender (Female)</td>
<td>+</td>
<td></td>
<td>+</td>
<td>★</td>
</tr>
<tr>
<td>log source length</td>
<td>+</td>
<td>★★★</td>
<td>+</td>
<td>★★★</td>
</tr>
<tr>
<td>ui (PTM) : ui order</td>
<td>+</td>
<td></td>
<td>−</td>
<td>★</td>
</tr>
</tbody>
</table>

The key independent variable: translation condition
Learning effect – People get quicker as the task proceed
More edits means longer time
Initial translation quality – how much edit is necessary
They had unbalanced participation pool
Longer source sentence takes longer to edit
Potential interaction between independent variables
+random intercepts/slopes for subject, source, text genre.
RQ1: Time - PTM faster than post-edit?

Likert Scale survey (can still quantitatively compare users’ subjective judgements):
“In which interface did you feel most productive?”
“I would use interactive translation features if they were integrated into ”
“I got better at using the interactive interface with practice/experience”

Think aloud (record users’ comments) – Interactive mode takes more time because…

There are more aids to operate and more information to read and analyze:
“Because you spend more time on each word, you have opportunity to see alternative translations.”

MT quality greatly affected the usefulness of the interactive aids:
“If drop-down suggestions are not of a good quality, reading (without selecting them) may consume extra time.”

The post-edit mode was easier at first, but in the end the interactive mode was better once I got used to it.
“I am used to this [post-edit], this is how Trados [the preeminent CAT tool] works.”
RQ2: Quality - PTM == better translation?

Metric: BLEU (automatic eval, has issues, but easier to run)
BLEU: a measure of similarity with the gold reference.
HBLEU: measure of similarity with the initial MT suggestions.

Compare mean
(Also vs. original generated text)

<table>
<thead>
<tr>
<th></th>
<th>Fr-En</th>
<th></th>
<th>En-De</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>HBLEU</td>
<td>BLEU</td>
<td>HBLEU</td>
</tr>
<tr>
<td>post-edit</td>
<td>38.1</td>
<td>63.7</td>
<td>29.4</td>
<td>44.1</td>
</tr>
<tr>
<td>PTM</td>
<td>38.4</td>
<td>62.6</td>
<td>29.5</td>
<td>41.0</td>
</tr>
</tbody>
</table>

“PTM exposes translators to many more alternatives, encouraging them to deviate further from the initial MT suggestion (lower HBLEU).”
RQ2: Quality - PTM == better translation?

Metric: Human subject rating

Auto methods are sensitive and noisy, so usually paired with human judgements as well.

![Image showing translation quality for En-De and Fr-En pairs with Human subject rating values]
RQ2: Quality - PTM == better translation?

Metric: BLEU & rating (automatic eval, has issues, but easier to run)
Qualitative analysis (find reasons behind quantitative analysis)

Why do many participants prefer post-edit?

“I found the machine translations (texts in gray) were of a much better quality than texts generated by Google Translate”
“The translations generally did not need too much editing, which is not always the case with machine translations.”

When users wanted to render more stylistic translations, PTM was less useful:

“...choosing a very different translation approach (choice of words, idioms with no equivalent in English...) would be like going against the current—but may have provided a better quality.”
“the translator is less susceptible to be creative.”
Conduct the Think Aloud: Test / Pilot the study

- Discover problems with study or concept being tested
- Estimate time needed for test
- Refine test script and tasks
- Verify typical tasks (something users actually do?)
- Practice before going live
- New ideas for follow-on questions or things to observe
Takeaways

User-centered design is **important**.

**Double Dimond** is a typical process.

**Reframing** the problem and the persona changes human behaviors.

**Interviews & think-aloud** are important HCI methods for building NLP-infused applications.

**Quantitative & qualitative** studies are both important.
Terminologies

Human Centered Design
User Centered Design
Value Centered Design
Humanity Centered Design
The Five Principles of Humanity-Centered Design

Humanity-centered design has five fundamental principles.