Crowdsourcing and Human Computer Interaction Design

Crowdsourcing and Human Computation
Lecture 10

Instructor: Chris Callison-Burch
TA: Ellie Pavlick

Website: crowdsourcing-class.org

Wizard of Oz in HCI

Wizard of Oz in HCI

Oz-like HCI in SciFi

AI is lacking compared to human intelligence. Some people earn a living as "ractors", interacting with customers in virtual reality entertainments. Ractors are more expensive than AI, so the only reason to use them is because customers can tell the difference. Virtual reality entertainment has become one ongoing Turing Test, and software is continuously failing it.

Wizard of Turk?

- Can we make SciFi a reality with crowdsourcing?
- Last week we examined the possibility of using humans as a function call in TurKit
- Can we use people in next generation interfaces for computers and mobile devices?
- What challenges does that present?

Word Processing: Boring HCI?

- Word processing supports a complex cognitive activity
- Writing is difficult: even experts routinely make style, grammar and spelling mistakes.
- Decisions like changing from past to present tense, or cutting 1/2 a page require many transformations across a document
- Current software provides little support for such tasks

Soylent: A Word Processor with a Crowd Inside

- Use large crowd of editors ala Wikipedia to improve your own work
- Use people's basic knowledge of English to edit the document to fix errors
- Opens up many other possibilities:
    - scan for superfluous words to trim
    - update addresses with zip codes
    - do things that Word cannot (false positives in spell check)

Soylent: A Word Processor with a Crowd Inside

- Implemented as a plugin to Microsoft Word using Microsoft Visual Studio Tools for Office (VSTO)
- Makes calls to Amazon Mechanical Turk with TurKit
- Has a set of 3 special purpose modules designed for work processing
    - Shortn
    - CrowdProof
    - The Human Macro

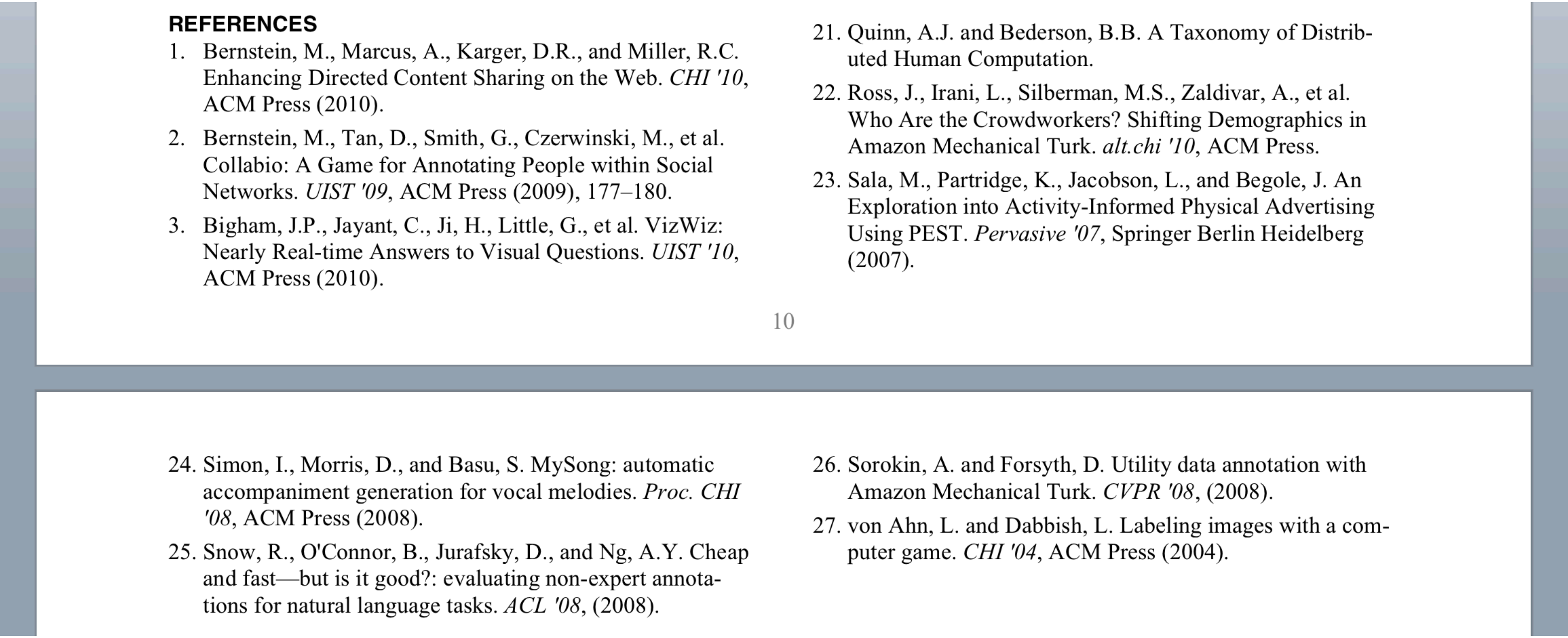
Shortn

- A text shortening service that cuts selected text down to 85% of its original length typically without changing the meaning of the text or introducing errors.

(Aside: Motivation for compression)

- Tweets are 140 characters
- Short URLs are ~20 characters
- Image descriptions target ~120 characters

Shortening a paper to 10 pages



REFERENCES

1. Bernstein, M., Marcus, A., Karger, D.R., and Miller, R.C. Enhancing Directed Content Sharing on the Web. *CHI '10*, ACM Press (2010).
2. Bernstein, M., Tan, D., Smith, G., Czerwinski, M., et al. Collabio: A Game for Annotating People within Social Networks. *UIST '09*, ACM Press (2009), 177–180.
3. Bigham, J.P., Jayant, C., Ji, H., Little, G., et al. VizWiz: Nearly Real-time Answers to Visual Questions. *UIST '10*, ACM Press (2010).

21. Quinn, A.J. and Bederson, B.B. A Taxonomy of Distributed Human Computation.
22. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., et al. Who Are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk. *alt.chi '10*, ACM Press.
23. Sala, M., Partridge, K., Jacobson, L., and Begole, J. An Exploration into Activity-Informed Physical Advertising Using PEST. *Pervasive '07*, Springer Berlin Heidelberg (2007).

24. Simon, I., Morris, D., and Basu, S. MySong: automatic accompaniment generation for vocal melodies. *Proc. CHI '08*, ACM Press (2008).
25. Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *ACL '08*, (2008).
26. Sorokin, A. and Forsyth, D. Utility data annotation with Amazon Mechanical Turk. *CVPR '08*, (2008).
27. von Ahn, L. and Dabbish, L. Labeling images with a computer game. *CHI '04*, ACM Press (2004).

AI approaches

- Rewriting text to be shorter is a task that Natural Language Processing researcher work on – including me and my students!
- The goal of “sentence compression” is to re-write text to be shorter while preserving all of its meaning

AI approaches

- Deletion
- Paraphrasing
- Summarization

AI approaches

Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.

AI approaches

🚫 Deletion

Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government~~, after days of haggling and tense hours of brinksmanship~~.

AI approaches

↔ Paraphrasing

Congress agreed Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.

Soylent’s solution

Congressional leaders reached a last-gasp agreement Friday to avert a shutdown of the federal government, after days of haggling and tense hours of brinksmanship.

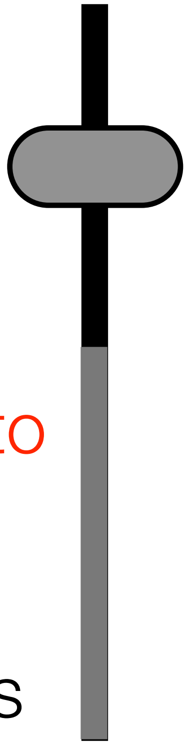
Shortn Interaction

- Selects the paragraph or section of text that is too long
- Press the Shortn button in the Word’s Soylent ribbon tab
- Soylent launches a series of MTurk Turk tasks and notifies user when text is ready
- User launches the Shortn dialog box

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.

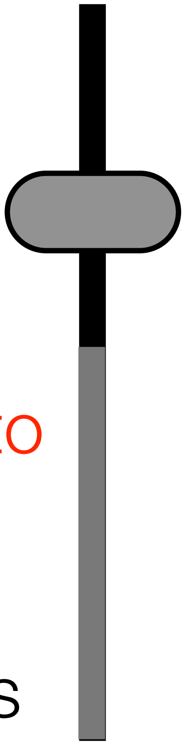
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.



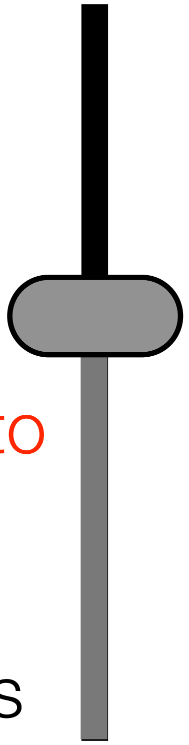
Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't relevant to a specific task. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, as structure differences aren't important to the editing task. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.

Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, because the differences in structure aren't important to the user's particular editing task. For example, if the user only needs to edit near the end of each line, then differences at the start of the line are largely irrelevant, and it isn't necessary to split base on those differences. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually, perhaps using drag-and-drop to merge and split clusters. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.



Automatic clustering generally helps separate different kinds of records that need to be edited differently, but it isn't perfect. Sometimes it creates more clusters than needed, as structure differences aren't important to the editing task. Conversely, sometimes the clustering isn't fine enough, leaving heterogeneous clusters that must be edited one line at a time. One solution to this problem would be to let the user rearrange the clustering manually using drag-and-drop edits. Clustering and selection generalization would also be improved by recognizing common test structure like URLs, filenames, email addresses, dates, times, etc.

Length reduction

- Reductions affect different parts of the text, so moving slider changes different regions
- Removes ~15–30% in a single pass, and up to ~50% with multiple iterations
- The algorithm preserves meaning, cutting only unnecessary language and repetitions
- User (not Workers) must remove whole arguments or sections

Example Shortn: Blog

Print publishers are in a tizzy over Apple's new iPad because they hope to ~~finally~~ be able to charge for their digital editions. But in order to get people to pay for their magazine and newspaper apps, they ~~are going to~~ have to offer something different that readers cannot get at the newsstand or on the open Web.

3 paragraphs, 12 sentences, 272 words	Reduced to 83% length of original	$4.57 187 workers	46–57 mins per paragraph

Shortn: Academic paper

The metaDESK effort is part of the larger Tangible Bits project~~. The Tangible Bits vision paper~~, which introduced the metaDESK ~~along with~~ and two companion platforms, the transBOARD and ambientROOM.

7 paragraphs 22 sentences 478 words	Reduced to 87% length of original	$7.45 264 workers	49–84 min per paragraph

Shortn: Academic paper

~~In this paper we argue that~~ it is possible and desirable to combine the easy input affordances of text with the powerful retrieval and visualization capabilities of graphical applications. We present WenSo, ~~a tool that~~ which uses lightweight text input to capture richly structured information for later retrieval and navigation in a graphical environment.

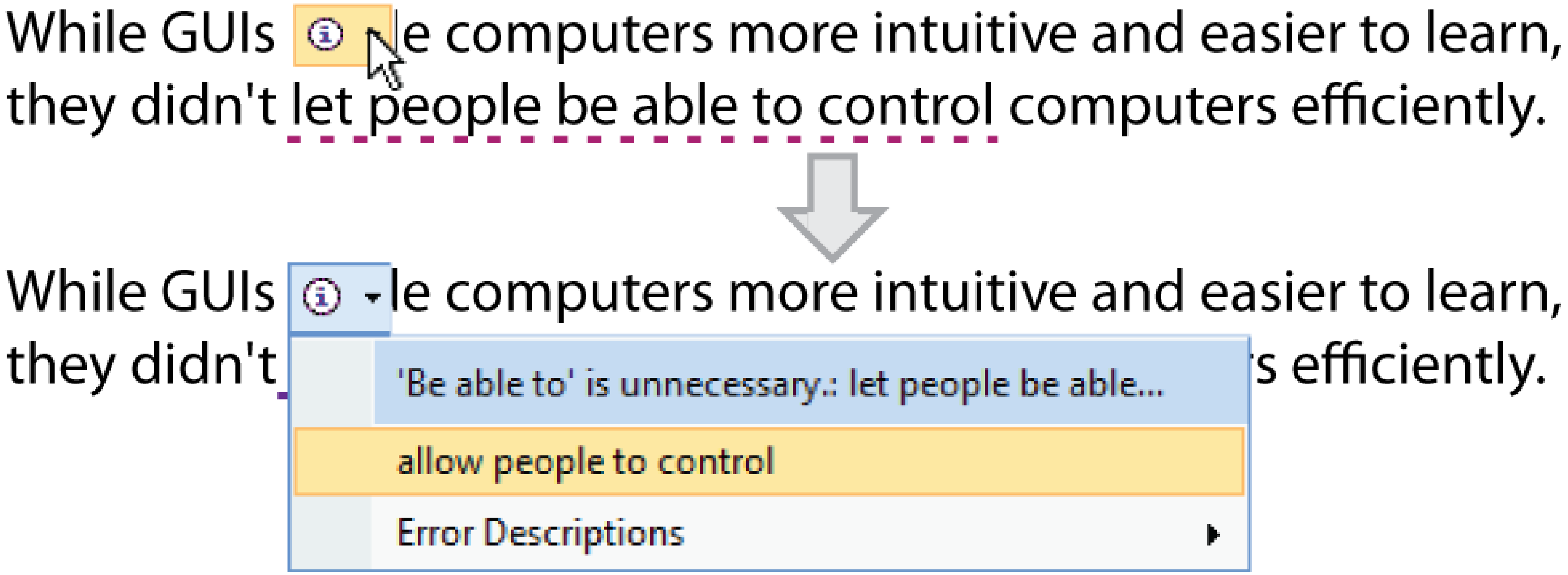
5 paragraphs 23 sentences 652 words	Reduced to 90% length of original	$7.47 284 workers	52–72 min per paragraph

Shortn: technical writing

Figure 3 shows the pseudocode that implements this design for Lookup. FAWN-DS extracts two fields from the 160-bit key: ~~the i low order bits of the key~~ (the index bits) and the next 15 low order bits ~~(the key fragment)~~.

3 paragraphs 13 sentences 291 words	Reduced to 82% length of original	$4.84 188 workers	132–489 min per paragraph

CrowdProof



While GUIs le computers more intuitive and easier to learn, they didn't let people be able to control computers efficiently.

While GUIs le computers more intuitive and easier to learn, they didn't s efficiently.

- A human-powered spelling and grammar checker that finds problems Word misses, explains the problems, and suggests fixes

Challenges for Soylent?

- In Soylent, Turkers are directly editing your documents
- What are the major concerns when other people are editing your documents?

High variance in user contributions

- Lazy workers – some workers do as little work as necessary to get paid
- Eager beavers – some do too much work or give random things that we didn’t ask for

Lazy worker

The theme of loneliness features throughout many scenes in Of Mice and Men and is often the dominant theme of sections during this story. This theme occurs during many circumstances but is not present from start to finish. In my mind for a theme to be pervasive is must be present during every element of the story. There are many themes that are present most of the way through such as sacrifice, friendship and comradeship. But in my opinion there is only one theme that is present from beginning to end, this theme is pursuit of dreams.

Eager Beaver

The theme of loneliness features throughout many scenes in Of Mice and Men and is often the principal, significant, primary, preeminent, prevailing, foremost, essential, crucial, vital, critical theme of sections during this story.

QC is hard

Insurance **company** may use the information to raise rates or to deny the insurance.
Insurance **company** may use the information to raise rates or to deny the insurance.
Insurance **company** may use the information to raise rates or to deny the insurance.
Insurance **companies** may use the information to raise rates or to deny the insurance.

Original	For **serendipity** discovery, the time taken is considered short.
Gold	For **serendipitous** discovery, the time taken is considered short.
distance = 33	**Serendipitous discoveries do not take long.**
distance = 3	**For serendipity discovery, the time taken is considered short.**

The find-fix-verify pattern

- No clear way to embed gold standard control data into tasks of this type
- Find-fix-verify is a 3 step process to try to ensure higher quality results
- Meant to correct the imbalance of work between lazy workers and eager beavers, and to reduce introduction of errors

Step 1: Find

- Identify passages that need improvement
- For proofreading: find at least 1 phrase or sentence that needs to be edited
- Aggregate across many independent opinions
- Regions with agreement are more likely to be correctable

Step 2: Fix

- Send the selected regions to other Worker to correct
- Each task now consists of a constrained edit to an area of interest
- Workers can see the whole paragraph but only edit the selected region
- 3-5 workers suggest alternate edits

Step 3: Verify

- Verify is a mechanism for performing quality control on the suggested edits
- Randomize the order of the proposed changes, and ask other Turkers to vote on the best one, or to flag poor suggestions
- Exclude workers who proposed the fixes, so they can't vote on their own work

Why use find-fix-verify?

- Why should tasks be split into independent Find-Fix-Verify stages?
- Why not let Turkers fix errors they find?
- Wouldn’t that be more efficient and cost effective?
- Does it solve problems with lazy workers? How?

Cost of find-fix-verify

	Shortn	Crowdproof
Find	$0.55	$0.06
Fix	$0.48	$0.08
Verify	$0.38	$0.04
Total	$1.41	$0.18
	per paragraph	per error

Crowdproof: ESL

However, while GUI made using computers ~~be~~ more intuitive and easier to learn, it didn’t ~~let people be able to~~ control computers efficiently. ~~Massesnis only can~~ The masses only can use the software developed by software companies, unless they know how to write programs.

1 paragraph 8 sentences 166 words	Errors caught: 5/12	$2.26 38 workers	47 minutes

Crowdproof: Notes

~~Blah blah blah —~~This is an argument about whether there should be a standard “~~nosql~~ NoSQL storage” API to protect developers storing their stuff in proprietary services in the cloud. ~~Probably unrealistic.~~ To protect yourself, use an open software offering, ~~and~~ self-host or go with hosting solution that uses open offering.

2 paragraphs 8 sentences 107 word	Errors caught: 8/14	$4.72 79 workers	42–53 minutes

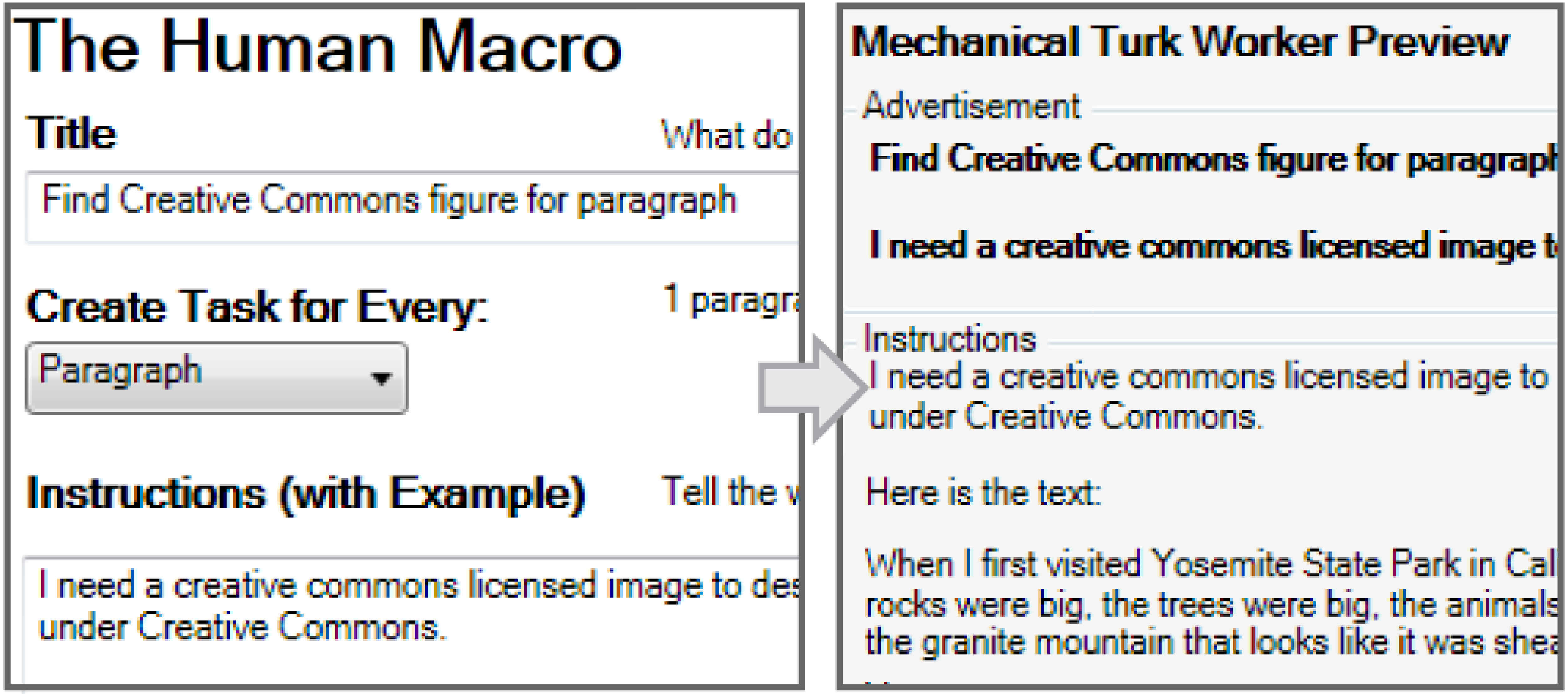
The Human Macro

- Macros usually require users to translate their intentions into algorithms explicitly via a scripting language
- The human macro is a “Natural Language Crowd Scripting Language”
- It allows the user to ask other people complete tasks like formatting citations or finding appropriate figures

Like Siri but unrestricted

- Natural language interfaces still struggle with unconstrained input
- Humans are good at understanding written instructions

The Human Macro



The Human Macro

Title

What do

Find Creative Commons figure for paragraph

Create Task for Every:

1 paragra

Paragraph

Instructions (with Example)

Tell the w

I need a creative commons licensed image to des
under Creative Commons.

Mechanical Turk Worker Preview

Advertisement

Find Creative Commons figure for paragraph

I need a creative commons licensed image t

Instructions

I need a creative commons licensed image to
under Creative Commons.

Here is the text:

When I first visited Yosemite State Park in Cal
rocks were big, the trees were big, the animals
the granite mountain that looks like it was shea

Design challenges

- Ensure that the user creates tasks that are scoped correctly for a Mechanical Turk worker
    - Ask user provide an example input and output, to clarify task requirements
- Prevent the user from spending money on a buggy command
    - The Human Macro helps debug the task by allowing a test run on a sentence or paragraph

Showing the results

- User specifies if Turkers’ work should replace the existing text or just annotate it
- If replace, text is underlined with drop-down substitution
- If annotate, feedback is inserted in comment bubbles anchored to selected text using Word’s comments interface

Human Macro Examples

Request	“Please change text in document from past tense to present tense.”
Input	I **gave** one final glance around before descending from the barrow. As I **did** so, my eye **caught** something [...]
Output	I **give** one final glance around before descending from the barrow. As I **do** so, my eye **catches** something [...]

Human Macro Examples

Request	“Pick out keywords from the paragraph like Yosemite, rock, half dome, park. Go to a site which has CC licensed images [...]”
Input	When I first visited Yosemite State Park in California, I was a boy. I was amazed by how big everything was [...]
Output	

Human Macro Examples

Request	“Please find the bibtex references for the 3 papers in brackets. You can located these by Google Scholar searches and clicking on bibtex.”
Input	Duncan and Watts [Duncan and watts HCOMP 09 anchoring] found that Turkers will do more work when you pay more, but that the quality is no higher.
Output	@conference{ title={{Financial incentives and [...]}}, author={Mason, W. and Watts, D.J.}, booktitle={HCOMP ‘09}}

Human Macro Examples

Request	“Please complete the addresses below to include all informtion needed as in example below. [...]”
Input	Max Marcus, 3416 colfax ave east, 80206
Output	Max Marcus 3416 E Colfax Ave Denver, CO 80206

Soylent’s contributions

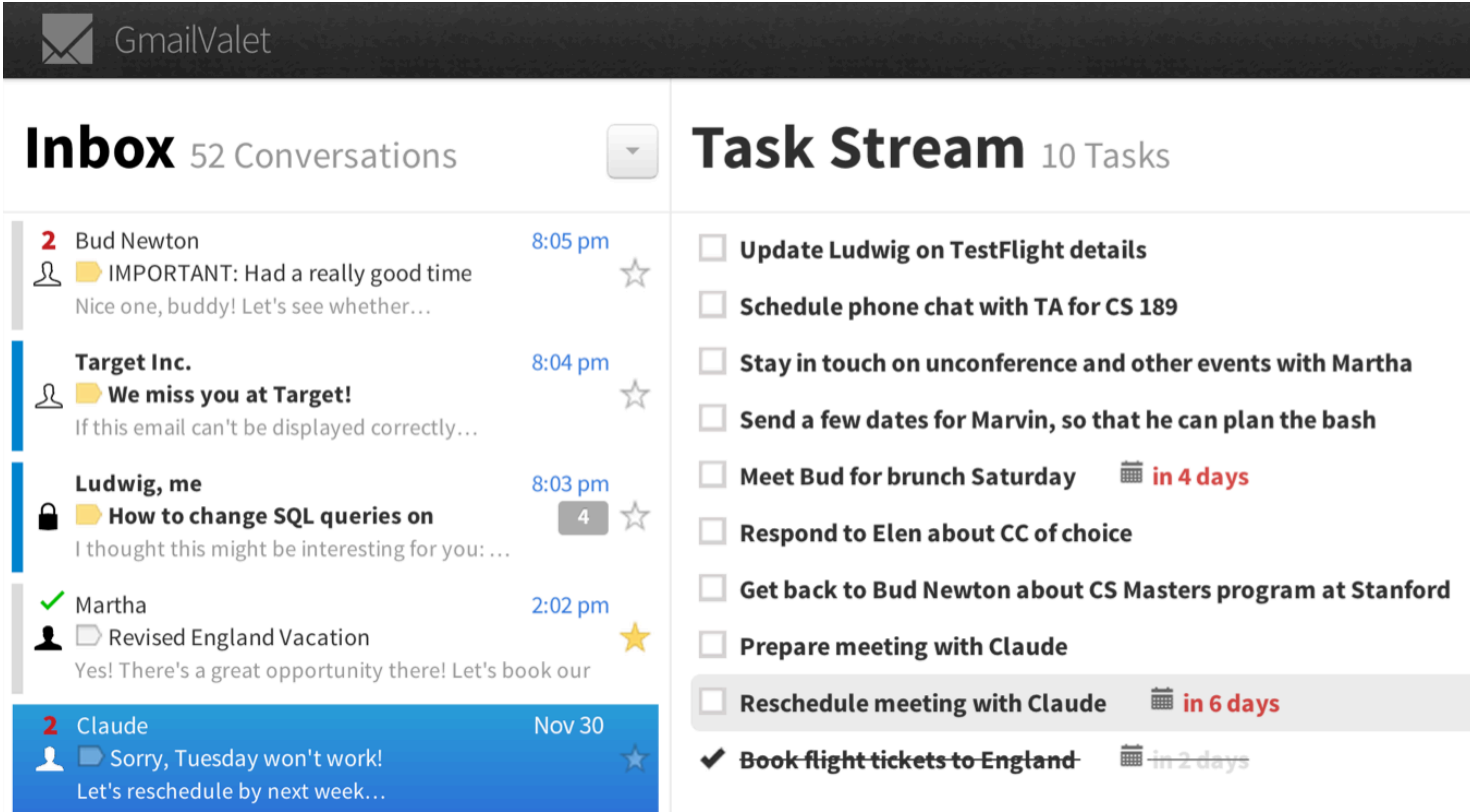
- The idea of embedding paid crowd workers in an interactive user interface to support complex cognition and manipulation tasks on demand
- Crowd workers can do HCI tasks that computers cannot reliably do automatically
- Easier to ask workers to do something than it is to write macro script

This paper presents Soylent, a word processing interface that uses crowd workers to help with proofreading, document shortening, editing and commenting tasks. Soylent is ~~an example of~~ a new ~~kind of~~ interactive user interface in which the end user has direct access to a crowd of workers for assistance with tasks that require human attention and common sense. Implementing these ~~kinds of~~ interfaces requires new software programming patterns ~~for interface software~~, since crowds behave differently than computer systems. We have introduced one important pattern, FindFix-Verify, which splits complex editing tasks into a series of identification, generation, and verification stages ~~that use independent agreement and voting~~ to produce reliable results. We evaluated Soylent with a range of editing tasks, finding and correcting 82% of grammar errors ~~when combined with automatic checking~~, shortening text to approximately 85% of original length per iteration, and executing a variety of human macros successfully.

Would you let just anyone edit your documents?

- Quality – do you believe that they are doing what we ask?
- Accuracy – do we have safeguards in place to avoid workers introducing errors?
- Privacy – do we trust them with the material? Is it sensitive?

Would you let them read your email?



GmailValet

Inbox 52 Conversations

2 Bud Newton 8:05 pm
IMPORTANT: Had a really good time
Nice one, buddy! Let's see whether…

Target Inc. 8:04 pm
We miss you at Target!
If this email can't be displayed correctly…

Ludwig, me 8:03 pm
How to change SQL queries on 4
I thought this might be interesting for you: …

✓ Martha 2:02 pm
Revised England Vacation
Yes! There's a great opportunity there! Let's book our

2 Claude Nov 30
Sorry, Tuesday won't work!
Let's reschedule by next week…

Task Stream 10 Tasks

- □ **Update Ludwig on TestFlight details**
- □ **Schedule phone chat with TA for CS 189**
- □ **Stay in touch on unconference and other events with Martha**
- □ **Send a few dates for Marvin, so that he can plan the bash**
- □ **Meet Bud for brunch Saturday** in 4 days
- □ **Respond to Elen about CC of choice**
- □ **Get back to Bud Newton about CS Masters program at Stanford**
- □ **Prepare meeting with Claude**
- □ **Reschedule meeting with Claude** in 6 days
- ✔ ~~**Book flight tickets to England**~~ ~~in 2 days~~