

Requirements

Ehud Reiter

What do users want?

- Must know what users want when build/eval NLG system!
 - *Requirements*
- Otherwise system (and evaluation) will be useless

Example: generating medical reports

- Users say accuracy is super-important
- So LLM which generates fluent-but-wrong texts is useless
- Eval which focuses on readability also not very useful

Software Requirements

- Getting requirements right important for ***all*** software
- Most big IT disasters due (at least in part) to wrong requirements
- NHS Connecting for Health
 - £12B electronic health record system
 - Good job of meeting *manager* requirements
 - Poor job of meeting *doctor/clinical* requirements
 - Never used
- I've seen several commercial NLG projects fail because of poor understanding of requirements

Discussion

- What are requirements for an NLG system that reports election results?
 - Input is number of votes per candidate in constituency
 - Output is news/media story about result
 - <https://www.bbc.co.uk/news/technology-50779761> (simple example)

Contents

- *Quality criteria*
- Workflow
- Acquiring requirements

Quality Criteria

- What aspects of a generated text do users care about?
- Readability
- Accuracy
- Content
- Utility
- Safety
- *Many more!*

Readability

- Texts should be easy/quick to read
 - For target users
- Example of poor readability (sports story)

Markieff Morris also had a nice game off the bench, as he scored 20 points and swatted away late in the fourth quarter to give the Suns a commanding Game 1 loss to give the Suns a 118-0 record in the Eastern Conference's first playoff series with at least the Eastern Conference win in Game 5.

Readability

- Entertainment (eg, fiction): very important
- News/media articles: quite important
- Professional (eg medical, legal, finance)
 - Essential that text be understandable and have decent readability
 - Less important that text be very easy to read

Accuracy

- The information communicated in a text should be true
 - With respect to real-world, input data, or both
 - Inferential accuracy as well as literal accuracy
- Example of poor accuracy (health)
 - GPT4 response to question about cancer support groups (CLAN is fine, URL is for spam site)

CLAN Cancer Support: They offer a range of services including counselling, relaxation therapies, and support groups.

- Address: CLAN House, 120 Westburn Road, Aberdeen, AB25 2QA.
- Website: [CLAN Cancer Support](<http://www.clanhouse.org>)

Accuracy

- Medicine, law, finance: very important
 - Sometimes distinguish between critical and non-critical errors
 - Critical error: *Patient is vomiting*
 - Non-critical error: *Patient's wife is vomiting*
 - Critical errors not acceptable, may tolerate a few non-critical
- News/media: important, but in practice may tolerate a small number of errors
- Fiction: not important

Content

- The text should communicate key insights to the user
 - I.e., key information should not be omitted
 - Different users may need different insights
- Example of missing content (health)
 - Summary of doctor-patient consultation does not say that patient is vomiting (when he is vomiting)
 - Summary says patient not vomiting: accuracy error
 - Summary says nothing about vomiting: content error

Content


- Medicine, law, finance: very important
 - As with accuracy, may distinguish between critical/non-critical omissions
- News/media: mixed, in part because unclear which insights are really important for users
- Fiction: not important

Utility









- Does the generated text help a user make a decision or perform a task?
- Example: ice-cream van (weather)

Example (from Dmitra Gkatzia)

Which shifts are least likely to have rain?

 Brad wants to only work 3 shifts tomorrow (he wants to go scuba diving).
He doesn't sell anything when it rains, so pick the three shifts where it is least likely to rain.

Pick the 3 shifts where it is least likely to rain

Shift 1	Shift 2	Shift 3	Shift 4
 light rain showers Chance of any rain 70% 	 sunny intervals Chance of any rain 30% 	 sun Chance of any rain 10% 	 sun Chance of any rain 0% 
Light rain showers are likely.	Sunny intervals with rain being possible - less likely than not.	Sunny with rain being unlikely.	Sunny with rain being extremely unlikely.
<input type="checkbox"/> Least likely to rain	<input checked="" type="checkbox"/> Least likely to rain	<input checked="" type="checkbox"/> Least likely to rain	<input checked="" type="checkbox"/> Least likely to rain

D Gkatzia et al (2016). Natural Language Generation enhances human decision-making with uncertain information. *Proc of ACL*

Utility

- In principle, very important in professional contexts
 - In practice, may be hard to define and measure
- May depend on non-technical factors
 - Hospital refused to allow us to put health information app on web, which made it much less useful

Safety

- Generated texts should not upset, mislead, or otherwise harm users (or third parties)
 - Texts should not show racial bias, use profanity, etc
- Example (health)

User: I have depression and anxiety disorder so I'm in treatment. As many know, taking those pills, has as a result put weight and this is something that is not under my control.

ChatGPT: It could be helpful to keep track of what you eat and your physical activity in a journal to identify patterns and make adjustments.

Safety

- Usually important
 - Governments increasingly require this (EU AI Act)
- Differs for different people
 - Cultural, racial, etc background

Average vs worst case

- Users may only care about *average* readability, accuracy, etc
 - On average texts are very readable
- Other times, users may want a guarantee that readability, etc of *all* texts will meet a minimum standard
 - All texts have decent readability
 - Constraint on worst-case readability (etc) of generated texts

Quality criteria for election results reporter

1. Safety: Essential!!
2. Accuracy: Important, but a few non-critical errors may be acceptable in practice
3. Readability: Important that texts have good readability. High readability is “nice to have”, not essential
4. Content: Less important, readers can consult other sources if they have special interests
5. Utility: Hard to define what this means here

Contents

- Quality criteria
- *Workflow*
- Acquiring requirements

Workflow

- Automatic NLG
- Human checks and edits NLG output
- NLG helps human

Automatic NLG

- NLG texts are sent to end users
 - No interaction with human authors
- Consumer NLG: ChatGPT, Gemini, etc
- Most academic research systems

Human checking

- NLG system generates draft text
- Human author checks and fixes text
 - *Post-editing*
- Example: weather forecasts for offshore oil rigs
 - NLG text:* SW 20-25 backing SSW 28-33 by midday, then gradually increasing 34-39 by midnight.
 - Human-edited version:* SW 22-27 gradually increasing SSW 34-39.

Human checking

- Very common in professional contexts
 - Doctors check/edit NLG medical report
 - Journalists check/edit NLG article
- Fix mistakes
- Add background
 - Previous interaction with patients
 - Relevant world events

Human checking

- Needs to be integrated into human workflow
 - When does doctor check medical report?
- Should be supported by good UI
 - How does doctor check medical report?
- Editing time should be less than time to write from scratch!
 - Otherwise no point in NLG
 - NoteGenerator: only 9% time savings (edit vs write-from-scratch), not worth it.

NLG helps human writer

- Human is main author, NLG system is assistant
- NLG system generates potential content, human can copy-and-edit useful chunks into text
 - Financial reports: NLG generates 10 potential paragraphs, human uses 3 of these
- NLG system improves English by paraphrasing text
 - Non-native speakers find this useful
- Brainstorming tool

NLG helps human writer

- Lots of possibilities here!
- Hopefully will see more investigation of how humans and NLG can support each other

Workflow for election results reporter

- **Human checking** – Journalists quickly check texts for correctness, may in some cases add additional content for “colour” or human interest
- Automatic – not acceptable, journalist must check in order to ensure all texts are acceptable
- NLG helps human – not acceptable here because of time pressure (need to get 100s of texts out in a few hours). Could make sense if less time pressure

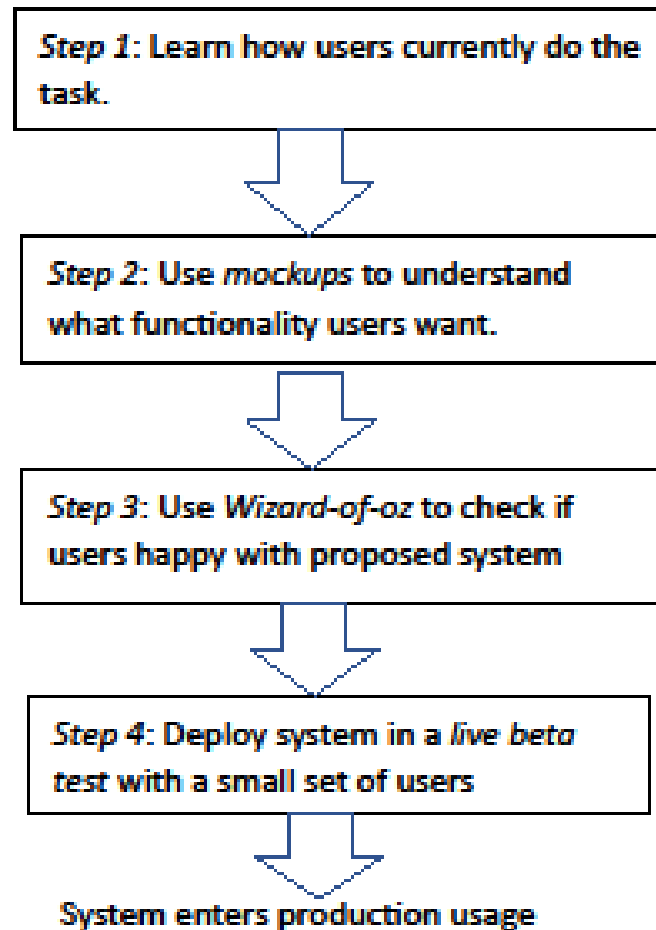
Contents

- Quality criteria
- Workflow
- *Acquiring requirements*

Requirements Acquisition

- Large literature on this in software engineering, much of which applies to NLG
- Techniques include user studies, mockups, Wizard of Oz, prototypes, beta test

From Knoll et al (2022)



T Knoll et al (2022). User-Driven Research of Medical Note Generation Software. *Proc of NAACL*

User studies

- If a task is currently done manually, find out how people do it
 - How do doctors currently write reports?
 - How do journalists currently write stories?
- Interviews, observational studies
- Objective
 - What information is needed to produce the text
 - Challenging/difficult cases
 - Where humans struggle and want help

Mockups

- Create powerpoint (etc) “mockup” of potential systems
 - Fake, not real!
 - Do this for several possible NLG systems
- Show mockups to users
 - Discard ones they dislike
 - Tweak ones they like to improve them
- Objective
 - High-level guidance on what users want

Prototypes or Wizard of Oz

- Create an initial version of system, and ask users to try it
 - Prototype: working code
 - Wizard of Oz: person pretends to be NLG system
- Refine based on user feedback
- Objective
 - More details on what users want
 - More understanding of edge cases

Beta test

- Ask a small number of switched-on users to use system for real
 - Monitor what happens
 - Identify problems, adjust requirements/functionality
- This is the original meaning of “beta test”
 - Some AI people use it to mean “don’t blame me if this breaks”

Req acquisition for election results reporter

- User studies – important, understand what journalists do
- Mockups – useful to create post-editing environment as well as refine core NLG system
- Prototypes – useful as part of debugging (journalists see problems which software devs don't notice)
- Beta test – not feasible (general elections only occur once every few years)

Conclusion

- Need to understand user requirements in order to build and evaluate NLG systems!
 - Which quality criteria matter most
 - Human-AI workflow
 - Also classic software requirements!
- Don't focus eval on readability if user cares more about accuracy
 - Even if readability is easier to measure!
 - I've seen a lot of researchers do this...