

Term Paper Outline

Cat Ball <catball@uw.edu>

2023-05-09 LING 575 A

Proposal paper:  LING 575 A (Bender) - Term Paper Proposal - Cat Ball

[0. Notes from meeting](#)

[1. Abstract](#)

[2. Introduction](#)

[2.1. Assumptions](#)

[2.2. Hypothesis](#)

[3. Background / Related Work](#)

[4. Methodology](#)

[4.1. Participants](#)

[4.1.1. Sample size](#)

[4.1.2. Participant demographics](#)

[4.1.3. Participant Recruitment](#)

[4.3 Participant Selection](#)

[4.4. Tasks](#)

[5. Results](#)

[5.1. Data Collection](#)

[5.2. Analysis](#)

[6. Discussion](#)

[7. Ethical Considerations](#)

[7.1. For Participants](#)

[7.2. For Everyone](#)

[8. Open Questions](#)

[9. Conclusion](#)

[10. Acknowledgements](#)

[11. Bibliography](#)

[12. Appendix](#)

1. Abstract

- This paper investigates conversational agents (e.g. chatbots) for IR tasks
- User study to see if people produce divulge more info about themselves when talking to a chatbot to do IR tasks
- Hypothesize people will produce entailments about themselves via implicature and presuppositions in the language they use with chatbots more frequently than they would with a query field

2. Introduction

This is a preregistration report (Nosek, 2018) for a user study on how people interact with information retrieval (IR) systems (specifically, internet search engines). The study aims to:

- Compare how people's language differ when they search for information via:
 - a text field (*e.g. Google, DuckDuckGo, etc*)
 - an artificial conversational agent (*e.g. ChatGPT for Bing, Bard for Google Search, etc*)
- Compare how much personal information is divulged in either scenario, and categories of personal info.
- Scope is limited to single-turn interaction
 - If hypothesis is confirmed, multi-turn interactions would be an interesting followup

2.1. Assumptions

1. The language people use to query IR systems is influenced by the user interface (UI)
2. Divulging personal information to strangers can have privacy and security implications

2.2. Hypothesis

- Compared to people using a text query field for IR, people using a conversational agent:
 - will be inclined to use more presuppositions and implicature in their query, and
 - their queries will entail more personal information about themselves (and possibly others)

3. Background / Related Work

- Preregistration format: (Nosek, 2018)
- Past work on IR & chatbots: (Shah, Bender 2022)
- Methodology: (Jokinen, 2010) and (Orne 2009)
- Presupposition triggers for identification: (Levinson 1983) and (Van der Sandt 1992)
- Implicature: (Grice 1975), (Levinson 2000)
- **todo:** find citations from previous ling department talk on user studies w/ race and ethnicity
talk about collecting minimizing PII data collection where not needed

4. Methodology

- The study will be an A/B test in the form of a survey
 - Reference (Kohavi 2023)
- Each question gives the participant an IR task (see section 4.4 Tasks below)
 - Tasks descriptions are the same for both groups
 - Jokinen 2010 talks a little on tasks, but would be nice if I can find some lit on IR-specific tasks
- Question then has an illustration of the UI they will be interacting with
 - Control group is shown a typical search engine text field UI
 - Experimental group is shown an instant messenger UI with a chatbot intro / welcome
- Survey will be written in English, and participants will be asked to write their responses in English
 - This is because I'm presently the only evaluator for data in this study, and I'm not very proficient in other languages, and do not want this to lead to inaccurate results because I misunderstood a response

4.1. Participants

4.1.1. *Sample size*

- Tradeoff between larger sample for better accuracy and smaller population to fit resource and time constraints posed by evaluation and analysis
- Diversity important for a representative sample; ideally multiple participants representing each desired demographic if possible

- After analysis, if sample is too small or is insufficiently diverse, recruit and select additional participants for additional rounds of the study (time permitting)

4.1.2. Participant demographics

- Demographic data will be collected via optional questions at the end of the survey
- Key demographics collected with Likert scale questions regarding:
 - Amount of past experience with:
 - search engines
 - conversational agents
 - Self-perceived computer proficiency
- Additional demographics will be collected to observe how different populations may be impacted
 - Age
 - Gender
 - Ethnicity
 - Self-perceived English proficiency
- It is possible that a participant may share additional demographic data inadvertently disclosed through free text fields in the survey while answering questions. Elicited demographic data should not be collected or linked to the participant.
- Initially I was going to skip collecting age, gender, and ethnicity with the idea that collecting less user data where it may not be necessary would be ideal for reducing risk of privacy breaches occurring, but talking to classmates, it sounds like this would be data worth gathering in case some groups are impacted disproportionately

4.1.3. Participant Recruitment

- Recruitment will include:
 - email solicitation to classmates and faculty at UW
 - email solicitation to coworkers
 - soliciting participants via social media
- The above are a convenient and cheap way to solicit volunteers, but may suffer from some selection bias
- Crowd workers (e.g. Amazon Mechanical Turk) may be solicited for additional data
 - Anecdotally, crowd-sourcing via AMT might yield quality issues

- Also it costs money
- If a budget for this appears, volunteers may be offered some form of compensation
- **todo:** figure out if there's resources / standard way to recruit participants at UW

4.3 Participant Selection

- Participants must be able to write survey responses in English
 - As the only evaluator, I'm most proficient in English and don't want to make a mess of the data by misunderstanding responses
- All interested may participate
 - If demographic groups are skewed, this should be noted during analysis of the data and normalized where appropriate

4.4. Tasks

- Study will be in the form of a survey
- The task is fictional scenario that they are unlikely to have biases toward, and without a "correct" answer to steer toward. (Orne, 2009)
- Tasks are followed by an image of the UI
 - Control group gets an image of a search query text box similar to major search engines.
 - Experimental group gets an image of a chat window with a conversational agent for search and a textbox to write in at the bottom.
 - Example text in the conversational agent may be something like *"Hello, I'm SearchBot. I'm an AI to help search for information on the internet. What can I help you with?"*
- Example questions
 - *"You recently acquired a pet Bipbo. It is three months old. You don't know what a Bipbo eats. What would you write in this search tool?"*
 - *"You heard that the nation of Bippopolis has elected a new president, but you don't know the new president's name. What would you write in this search tool?"*
- Solicit reviewers other linguists and HCI folks to review questions to reduce biasing participants in how questions are asked

5. Results

5.1. Data Collection

- Outline tool used to produce surveys (e.g. surveymonkey, typeform, etc)
- Any PII for participants (such as contact info, name, etc) should be stored separate from responses to reduce likelihood of correlating responses. Response table keyed on UUID
 - cite Jokeinen, McTear

5.2. Analysis

- Analyze entailments present in responses
- Key metrics:
 - Total entailments, grouped by question and participant group.
 - If the question proposes a new fact about the participant, does the participant's response entail the new fact?
 - Quantity and quality of additional entailments about the participant.
 - Quantity and quality of other entailments.
- Identifying presupposition and implicature and what they entail
 - Tools may be used in an initial pass to assist human evaluators, e.g. IMPPRES (Jeretic 2020)
 - For manual assessment, refer to
 - presupposition triggers from (Levinson 1983) and (Van der Sandt 1992)
 - (Levinson 2000) for implicature
- If more evaluators appear, cross-validate evaluations
- If participants' responses entail personal information about themselves, track how many instances where this occurs, and categorize the types of personal information divulged
 - Do not publish specific responses or quotes containing personal information about participants
 - Only do this if there is a non-trivial portion of responses with personal info to avoid the possibilities of someone eliding the identity of a participant
- Cluster metrics by demographic
 - This may help indicate how people from different backgrounds may be more or less inclined to insert entailments

6. Discussion

- If hypothesis is true, it may suggest additional privacy & security risk to users interfacing with IR systems via a chat interface with a conversational agent
 - Point to (Shah, Bender, 2020) that this is bad for a bunch more reasons too
- Weigh additional risks in conversational agents for IR against advantages of using conversational agents
 - There does not seem to be any compelling literature stating why someone might want a conversational agent for IR
- Frown loudly at search companies

7. Ethical Considerations

7.1. For Participants

- Be good stewards of data to ensure privacy of participants; see notes in section 5 on handling data appropriately
 - If participant responses were compromised, either unintentionally or maliciously, we want to reduce the risk of a participants being identified; hence keeping responses in a separate data store than PII, and reducing the amount of data collected

7.2. For Everyone

- If the hypothesis is true, this may allow those with access to the chatbot's logs to create an accurate model of those who interact with it (compared to traditional search query fields)
 - Likely use-case is advertising
- Privacy, security
 - Cite reports of people identifying individuals using pseudo-anonymous data from data brokers
 - Bad actor or bugs may leak sufficient user data to identify users

8. Open Questions

- Does IR with a convo agent usually produce a single-turn discourse per query, or multi-turn?
- Multi-turn conversation study, maybe Wizard of Oz experiment

- see: (Jokinen, 2010) pg. 102 for WoZ experiment methodology overview
- Are results found here similar for other languages?
- Investigate if any communities disproportionately impacted by this
- Does the amount of anthropomorphization applied to the conversation agent impact this?

9. Conclusion

- Study aims to understand how user's may disclose personal information, intentionally or unintentionally, to a conversational agent
 - via A/B testing
 - use fictional tasks & data handling best practices to

10. Acknowledgements

- Classmates
 - Elizabeth Okada & Hanieh Nezakati for reviewing outline
- Professor
 - Dr. Emily Bender for instructing LING 575 A and for advising and feedback on this work
- Authors
 - Thank authors of papers influential to this work
- Anyone else who shows up between now and this paper being done :)


11. Bibliography

- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.
- Kristiina Jokinen and Michael McTear. 2010. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESSive? Learning IMPlicature and PRESupposition. In

Proceedings of the 58th annual meeting of the association for computational linguistics, pages 8690–8705, Online. Association for Computational Linguistics.

- Stephen C. Levinson. 1983. *Pragmatics*, Cambridge Textbooks in Linguistics, pages 181–184, 194. Cambridge University Press
- Rob A. Van der Sandt. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics*, 9(4):333–377.
- Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.
- H. P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, pages 41–58. BRILL.
- Ron Kohavi and Roger Longbotham. 2023. Online Controlled Experiments and A/B Tests. In Dinh Phung, Geoffrey I. Webb, and Claude Sammut, editors, *Encyclopedia of Machine Learning and Data Science*, pages 1–13. Springer US, New York, NY.

12. Appendix

1. Term paper requirements:
 - a. https://faculty.washington.edu/ebender/2023_575/term-project.html
2. Term paper proposal
 - a.  LING 575 A (Bender) - Term Paper Proposal - Cat Ball