Experimental Design for CS Experiments

Ehud Reiter

University of Aberdeen

Experimental Design

When designing an experiment (for evaluation or otherwise), we need to decide on

- Hypotheses
- Subjects (*if people are involved*)
- Material
- Procedure
- Analysis [stats]

Get the basics right!

Important part of CS4040 project!

Example

- Assume we have invented a new search engine SuperSearch, which we think is better than Google
- How do we evaluate SuperSearch?
  - » Hypotheses
  - » Subjects
  - » Material
  - » Procedure
  - » Analysis

Hypothesis

- We need to refine qualitative “better than” into something we can measure

Possible Hypotheses

- User satisfaction:
  - » Users explicitly prefer SuperSearch
  - » Users give SuperSearch higher ratings
- Task performance
  - » Ask people to find some information, see if quicker (fewer mistakes) with SuperSearch
- Metrics
  - » Number of top-10 hits which are relevant

Subjects

- If we test on users, who are they?
- *Note*: only relevant if we use human subjects, many CS experiments are purely computational.

Possible Subjects

- If SuperSearch is for everyone, get diverse people from different background
  - » At least 100, 1000 would be better?
- Recruit via crowdsourcing?

Material

- What specific web searches (or tasks) do we use to evaluate SuperSearch?

Possible Material

- Most popular searches
  - » From Google Trends?
  - » Mostly companies and brands
- Challenging
  - » Complex and difficult searches?
- Random: Let subjects search for whatever interests them

Procedure

- What do we ask subjects to do?
- What do we compute and measure?

Possible Procedure

- Many possibilities!
- Simple example
  - » All subjects use both SuperSearch and Google, on all of our searches
  - » At end, we ask subjects which they preferred

Analysis

- How do we analyse and report the data?

Possible Analysis

- Depends on procedure, etc
- Simple procedure
  - » Report percentage of people who preferred SuperSearch, Google
  - » Test for statistical significance (*future lec*)
    - – Null hypothesis: 50-50 split in preference
    - – Binomial test

Eval SuperSearch

- One possible design
  - » Hypotheses: users prefer SS over Google
  - » Subjects: 1000 crowdworkers
  - » Material: subjects chose own queries
  - » Procedure: try queries on both SS and Google, see which preferred
  - » Analysis: Binomial test of preferences
- Many others!

More Details

- **Hypotheses**
- Subjects
- Material
- Procedure
- *Analysis*

Hyp vs Research Questions

- Sometimes distinguish between
- *Research Question*
  - » High-level, eg do users prefer SuperSearch over Google?
- *Hypotheses*
  - » Detailed, eg do users give a higher Likert rating to SuperSeach compared to Google when trying to find academic papers
- Ill use above interchangably

Hypotheses in Evaluation Experiments

- CS evaluation hypotheses are usually about
  - » *Utility*: System helps user do a task, eg find information
  - » *Performance (metrics):* System will have a good score on accuracy, precision, etc
  - » *User satisfaction*: Users will like and be satisfied by the system
  - » *Compute speed*: System will be fast
- Other possibilities, eg similarity to humans

Compare to baseline

- Usually hypothesis involves comparing system to existing “baseline” (eg Google)
  - » System more useful than existing “baseline”
    - – Also called “control”
  - » Users will prefer system over baseline
  - » System will have better accuracy than baseline

Which Baselines?

- Best performing existing system (alg)?
  - » Common in academic research
- Market leader?
  - » Common in commercial work
- Person doing this task?
  - » Occasionally done in complex tasks such as medical diagnosis

Weak baselines

- Choosing weak/inappropriate baselines is an easy/common way to “cheat!
  - » Compare against “state of the art” in 2025
  - » Ignoring simple non-neural baselines, like using first sentence of new article as a summary

Reminder: Hypoth Before Exper

- Decide on hypotheses ***before*** you do experiment
  - » Write them down somewhere
- Don’t change/tweak hypotheses to better fit data
  - » “Results weren’t good, so I played around with data until I found something significant”
  - » Common way to cheat

Experimental Design

- Hypotheses
- **Subjects**
- Material
- Procedure
- Analysis

Subjects

- If we're doing experiments on people, who are they?
- How many subjects are we looking for?
- How do we recruit them?

Who are Subjects?

- Subjects should be potential users
  - » If SuperSearch is intended for lawyers don’t ask CS students to evaluate it!
- Subjects should be representative
  - » age, gender, expertise, etc
    - –Can be hard to achieve in practice
  - » You should report subject characteristics
    - –Eg, students, ages18-25, 45% female

How Many Subjects?

- Ideally numbers based on a statistical power calculation
  - » Advanced stats lecture
- Hard to do in practice
- Student projects often use 20 subjects

Subject Recruitment

- Easiest recruitment is crowdsourcing, eg Amazon’s Mechanical Turk
  - » Amazon service where you can hire random people to do small tasks cheaply.
    - – “Task” is participating in your experiment
    - – Many alternatives to Amazon
- Works *if* Turkers are potential users, take task seriously, real-world context and domain knowledge not needed, and no need to observe or debrief subjects

Example

- Evaluate Chinese->English MT by showing output and a “reference translation” to an English speaker.
- Evaluate Chinese->English MT by showing Chinese input and English output to a Chinese->English human translator.
- Second is ***much*** better

Subject Recruitment

- Can recruit friends, colleagues, social network
    - » Free (unlike Amazon Mturk)
    - » Are they potential users, representative?
    - » Can you get enough subjects?
    - » Will they be biased because they know the outcome you hope to achieve?
        - –Know SuperSearch is your creation?

Subject Recruitment

- Explicit recruitment of subjects
  - » Via contacts, bulletin boards, advertisements, conferences, …
  - » Target exactly the kind of subject you want
  - » Often takes a lot of time and effort…

Experimental Design

- Hypotheses
- Subjects
- **Material**
- Procedure
- Analysis

Material

- Usually experiments are based on scenarios, which are defined by a set of input data
  - » Sometimes also include expected results
- How do we choose scenarios?
- How many scenarios do we need?

Standard data sets

- Sometimes we evaluate systems on standard data sets defined by someone else
  - » Common in machine learning
- If so, no need to worry about choosing scenarios

Material

- For search, scenario could consist of a search term and expected results
    - » Search: “Ehud Reiter’s home page”
    - » Expected results:
        - – https://www.abdn.ac.uk/ncs/profiles/e.reiter/
        - – https://ehudreiter.com/

Choosing Scenarios

- Common scenarios
  - » Searches in Google Trends
- Random scenarios
  - » Let users search whatever they want
- Difficult scenarios
  - » Technically challenging searches
  - » “most popular 2013 science fiction novel”

How do we choose scenarios?

- Often mixture
  - » Most common inputs
  - » Difficult inputs
  - » Random inputs

How Many Scenarios?

- Often determined by subject numbers and experimental desifn
  - » Eg, if 50 subjects each look at 4 scenarios, and we want each scenario seen by 10 subjects, then we should have 20 scenarios
- Otherwise, more is better!

Experimental Design

- Hypotheses
- Subjects
- Material
- **Procedure**
- Analysis

Procedure

- What do software do we use?
- What do we measure?
- What material does each subject see
  - » In what order
- Where does the experiment take place
- Etc

What software do we use?

- Quicksort
  - » Which programming language?
  - » Which implementation?
- Large language models
  - » Which version of GPT?
  - » Which prompt?
  - » Temperature?

What do we measure?

- From hypothesis
- But details need to be filled in
  - User satisfaction: 5-pt Likert scale or 7-pt Likert scale?
  - Compute speed: on what hardware?
  - Precision of search results: Top 10? Top 100? Top 1000?

Output quality

- Correctness against "gold standard"
  - » Sorting algorithm: Is output actually sorted?
  - » Face recognition: Is correct person IDed
  - » Can do automatically (without people)
- Assessment by human subjects
  - » Likert scale
  - » Error annotation

Likert Scale

- Present a statement to subjects, and ask if they agree on N-pt scale

  SuperSearch is better than Google

  1) Strongly disagree

  2) Disagreee

  3) Neither agree not disagree

  4) Agree

  5) Strongly agree

https://en.wikipedia.org/wiki/Likert_scale

Within/between subjects

- Within subjects: Subjects assess outputs of different sys on same scenario
  - » Sam checks Google and Bing on “Ehud Reiter’s home page”
  - » Tom checks Google and Bing on “Matthew Colliinson’s home page”
- Works well if possible
  - » Doesn’t make sense in some contexts

Within/between subjects

- Between subjects: Subjects assess outputs of different sys on diff scenario
  - Sam checks Google on “Ehud Reiter’s home page” and “Matthew Colliinson’s home page”
  - Tom checks Bing on “Ehud Reiter’s home page” and “Matthew Colliinson’s home page”

Latin Square

- Between-subjects design where every subject tries all systems an equal number of times
- Every scenario exposed an equal number of times in each system

Latin Square

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Subject 1	Google	SuperSearch	Google	SuperSearch
Subject 2	SuperSearch	Google	SuperSearch	Google
Subject 3	Google	SuperSearch	Google	SuperSearch
Subject 4	SuperSearch	Google	SuperSearch	Google

Procedure: Other

- *Practice scenarios*: Can give subjects some practice scenarios which we do not record
  - » Especially useful if we are timing people, since people are usually slower the first time
- *Exclusion*: May drop subjects who don’t seem to be taking the experiment seriously
  - » Drop outliers in general?
- *Ethics*

Experimental Design

When designing an experiment, we need to decide on

- Hypotheses
- Subjects
- Material
- Procedure
- **Analysis**

Analysis

We have the data, how do we analyse and report it?

- Show raw data (graphs, tables)
- Statistical analysis (*future lecture*)
- Error analysis

Error Analysis

- Find a few cases where the system failed, and try to understand why
  - Qualitative, not quantitative
- Face recognition gave wrong result because of poor lighting
- Machine translation failed to translate correctly when input was a poem

Experimental Design

- Very important, and key aspect of CS4040 report and honours project!

Ehud’s blogs

- I have written many blogs on exper des
- Challenges in Evaluating LLMs
  - » https://ehudreiter.com/2024/07/10/challenges-in-evaluating-llms/
- Ten tips on doing a good evaluation
  - » https://ehudreiter.com/2024/04/08/ten-tips-on-doing-a-good-evaluation/
- Common Flaws in NLP Evaluation Experiments
  - » https://ehudreiter.com/2024/01/15/common-flaws-in-nlp-evaluation-experiments/
- Many more!