Other Topics

Ehud Reiter

Contents

- *Commercial evaluation*
- Evaluation research
- Ehud’s advice and blog
- Discussion

Commercial Evaluation

- Companies and real-world users care about
  - Cost
  - Benefits
  - Risks
  - Return on investment

Cost

- What does it cost to develop an NLG system?
  - People
  - IT costs: GPU, cloud services
  - Marketing
  - Overheads
- What does it cost to maintain an NLG system?
- Students usually underestimate these!

Benefits

- What benefits come from the NLG system?
  - Software house: who will buy, how much will they pay?
  - Internal: productivity increases, cost savings, greater consistency, empowering junior/new staff, etc
- Benefits only come if people use the system!
  - Users may be hostile, especially if see as threat to their job
  - Change management

Risks

- Could developer be hit by massive lawsuits if something goes wrong?
  - Chatbot gives bad medical advice
  - Self-driving car hits pedestrian
  - Sensitive data stolen by hackers
  - Etc
- If above happens, who is liable?

Return on investment (ROI)

- How does investing in NLG compare to other investments
  - Maybe get more benefit from better training?
- Depends on alternative opportunities
- Also depends on time for NLG benefits to appear
  - Benefits in 1 year have better ROI than benefits in 5 years

Example: Medical report generator

- ***Note***: This is inspired by a real system I was involved with, but the numbers are made up (real numbers are commercial confidential)
- Suppose we have an NLG system that generates a summary of a consultation between a doctor and a patient
  - Doctors check and edit
  - Summaries must be written (legally required)

Medical Report Generator

- Cost: 10 person-years to develop, 2 person-year (annual) maintain
  - £1M development (including overheads)
  - £200K/year maintenance (including overheads)
- Benefits: Saves doctors 20 mins per day; used by 100 doctors
  - Checking-and-editing quicker than writing from scratch
  - £500K efficiency savings (doctors are expensive)
- Risk: Will mistake be made and not caught by doctors?
  - Judged low after extensive testing
- ROI: £1M up-front investment leads to £300K/year return
  - 30%, which is good

Scenarios

- What if development and maintenance are 2x higher?
  - Cost overruns common in software!
  - ROI drops to 5%, which is poor for new tech
- What is only 50 doctors use the system?
  - Rest are hostile or refuse to adapt to new workflows
  - ROI again drops to 5%
- What if system can only be used for 3 years
  - Changes in medicine and AI make it obsolete
  - ROI is negative (costs exceed benefits)

Commercial evaluation

- Need to estimate costs, benefits, risks, ROI
  - Central case and scenarios
- Often high uncertainty with new tech and workflows
  - Some companies keen on taking a risk
  - Most companies less keen

Contents

- Commercial evaluation
- *Evaluation research*
- Ehud’s advice and blog
- Discussion

Evaluation research

- Many researchers work on improving evaluation
- A *few* 2024 papers which I liked
  - Personal selection!

Experimental errors

C Thomson et al (2024). Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*.

https://doi.org/10.1162/coli_a_00508

- Replication studies reveal errors in executing evaluations
  - Buggy code, poor UI, numbers don't match data, etc
- Already mentioned in previous talk

Data contamination

S Ballocu et al (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. Proc of EACL 2024. https://aclanthology.org/2024.eacl-long.5/

- Structured survey of data contamination problems in recent NLP papers
- Shows could be an issue in ***42%*** of them!
  - Scary
- Hopefully people are becoming more sensitive to this

Metric validation: do small diff matter?

T Kocmi et al (2024). Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. Proc of ACL-2024

https://aclanthology.org/2024.acl-long.110/

- If system A gets a metric score of NA, which is higher than system B's metric score of NB, will users perceive A as better than B?
  - Eg, how big does the difference in score (NA-NB) need to be in order for users to agree with metrics assessment (which sys is better) 90% of time?
  - BLEU: never get 90% agree, even with huge diff in BLEU score
  - chrF: need diff of 3.05 points
  - Bleurt (default): need diff of 5.98 points
  - etc

Poor guidelines in human evaluations

J Ruan et al (2024). Defining and Detecting Vulnerability in Human Evaluation Guidelines: A Preliminary Study Towards Reliable NLG Evaluation. Proc of NAACL-2024

https://aclanthology.org/2024.naacl-long.441/

- Survey of human evaluations shows many (most?) do not give adequate guidelines to subjects
- Shows that LLMs can help detect problems and write better guidelines.

How valid are LLM-based evaluations?

A Bavaresco et al (2024). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. Arxiv

https://arxiv.org/abs/2406.18403

- Do LLM-as-judge correlate with good human evaluations?
  - Mixed
  - “This means that current LLMs and/or their prompts need to be calibrated against actual human judgments on every new dataset to establish the validity of their evaluation scores”
- Release dataset of assessing this

Questionable practices in machine learning

G Leech et al (2024). Questionable practices in machine learning.

https://arxiv.org/abs/2407.12220

- Discusses questionable practices in ML evaluation which lead to unreliable results.
- Expands on many of the concerns I have raised, adds new ones
  - Focus on contamination, cherrypicking, misreporting
- Don’t believe all published claims about amazing LLM perform!

Evaluation research

- Many more interesting papers!
- Great to see so much interest in evaluation

Contents

- Commercial evaluation
- Evaluation research
- *Ehud's advice and blog*
- Discussion

Ehud’s blog

- ehudreiter.com
- Regular blog about building and evaluating NLG systems
- Hopefully a good resource
- A few 2024 blogs which may be of interest

Building NLG sys: Latest tech may not be best

- When building an NLG system, it doesn't always make sense to use the latest technology.
- Understand the options and choose appropriate tech for use case

https://ehudreiter.com/2024/08/26/the-latest-trendiest-tech-isnt-always-appropriate/

Building NLG Sys: Adoption is slow in health

- Adoption of AI and NLG in healthcare is very slow
- NLG developers need to understand actual requirements and pain points of users

https://ehudreiter.com/2024/07/23/slow-adoption-of-ai-in-healthcare/

Evaluation: Challenges in evaluating LLM

- Evaluating modern LLMs is hard!
- Based on a workshop talk

https://ehudreiter.com/2024/07/10/challenges-in-evaluating-llms/

Evaluation: Subjects must understand task

- Biggest weakness in a lot of human evaluations is that subjects don’t understand what they are supposed to be doing
- Must properly explain to (and possibly train) subjects

https://ehudreiter.com/2024/05/28/human-eval-subjects-must-understand-the-task/

Evaluation: Ten tips

- Tips on doing good evaluation
- Largely covered in previous talks

https://ehudreiter.com/2024/04/08/ten-tips-on-doing-a-good-evaluation/

Academics: Systematic reviews

- Systematic reviews are a great way of bringing together evidence (including evaluation)
- Common practice in medicine
- Starting to become more popular in NLP

https://ehudreiter.com/2024/01/31/systematic-reviews-in-nlp/

Contents

- Commercial evaluation
- Evaluation research
- Ehud’s advice and blog
- *Discussion*