Towards the Reproducible Evaluation of Generative Information Retrieval Systems

Niklas Deckers and Martin Potthast
Leipzig University and ScaDS.AI

Workshop on the Impact of Generative AI on Search and Search Engine Research
Hamburg · September 7, 2023

Motivation

- Quality problems induced by the LLMs and the user often does not realize
- Models change quickly - making a reproducible and comparable evaluation difficult

Generative Models as an Index

- Inspired by the idea of the Infinite Index:

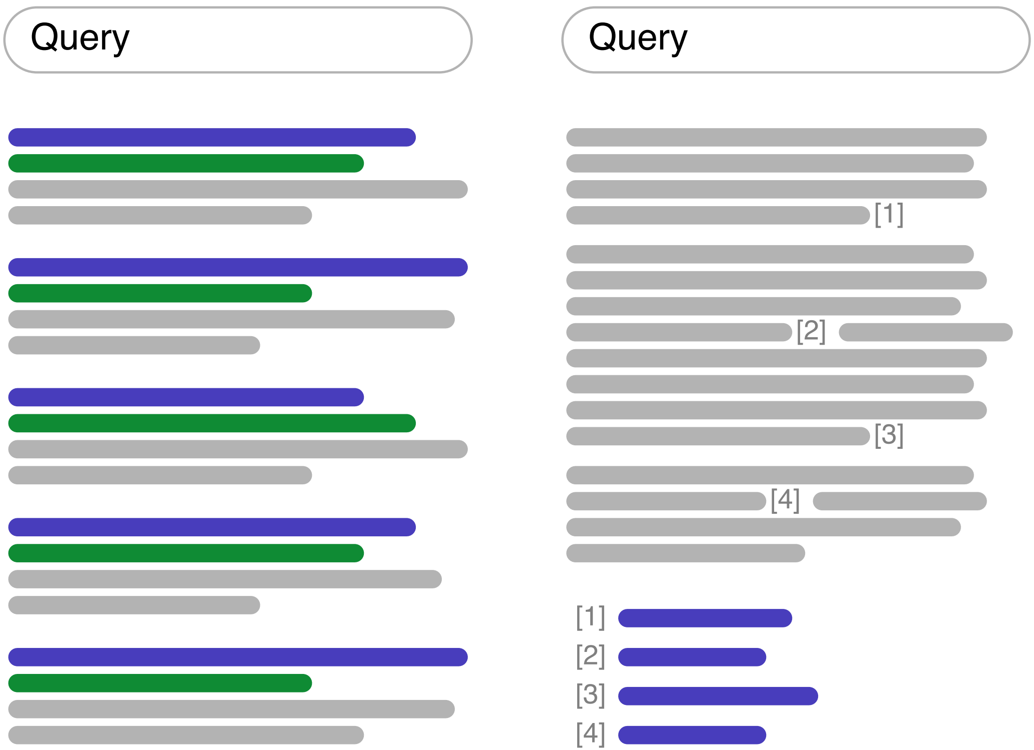
See generation with a prompt as
retrieval with a query, but on an infinite index

- Fundamental difference: Set of documents that is being retrieved on
- Will try to identify and use parallels between traditional and generative IR

Generative IR Systems

- Conversational approach
- Answering a question in natural language
- Including information and references from the web

List SERP vs. Text SERP



- SERPs are traditionally lists of document references (10 blue links)
- LLMs generate text documents with optional source references (text SERPs)

Components for the Evaluation of Generative IR Systems

① User Models

② Evaluation Metrics

③ Systems for Reproducible Evaluation Experiments

①

User Models

Applying the Accumulation Model

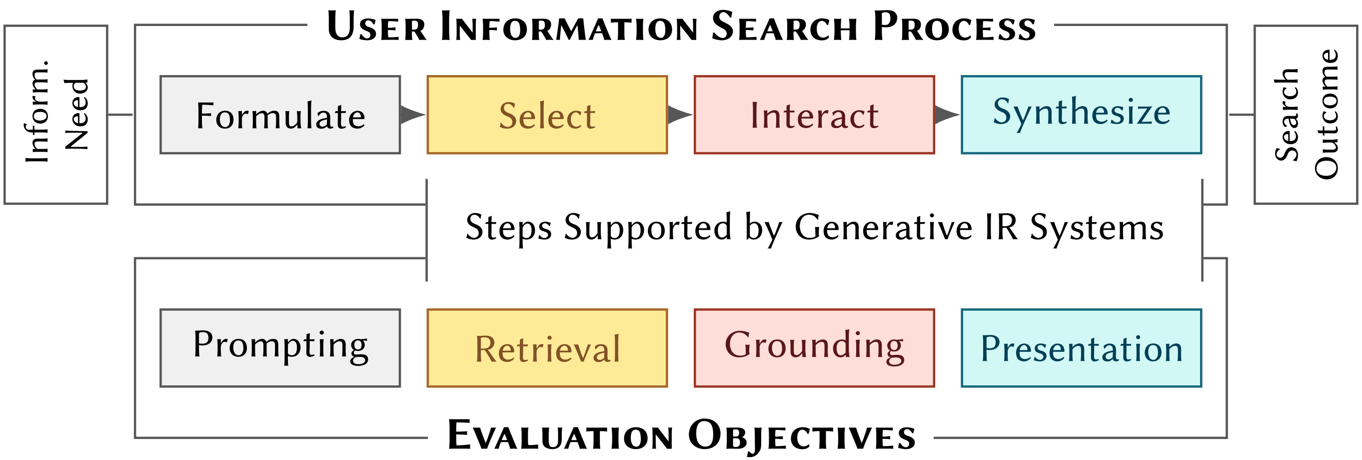
- Traditional IR:
  - a utility model (how each result provides utility to the user)
  - a browsing model (how the user interacts with results)
  - an accumulation model (how individual utility of documents is aggregated)
- Applying this idea to generative IR
- Evaluation will require segmentation into statements
- Results in a measure that looks similar to discounted cumulative gain (DCG)

②

Evaluation Metrics

Evaluation Objectives

- Evaluation objectives must be grounded in the underlying user model



Baselines

- Retrieval, then generation:
  - Provide LLM with query and snippets from a traditional IR system
  - Relies on high context length

- Generation, then retrieval:
  - Use LLM to generate a better query that is piped into a traditional IR system

- Approaches in between

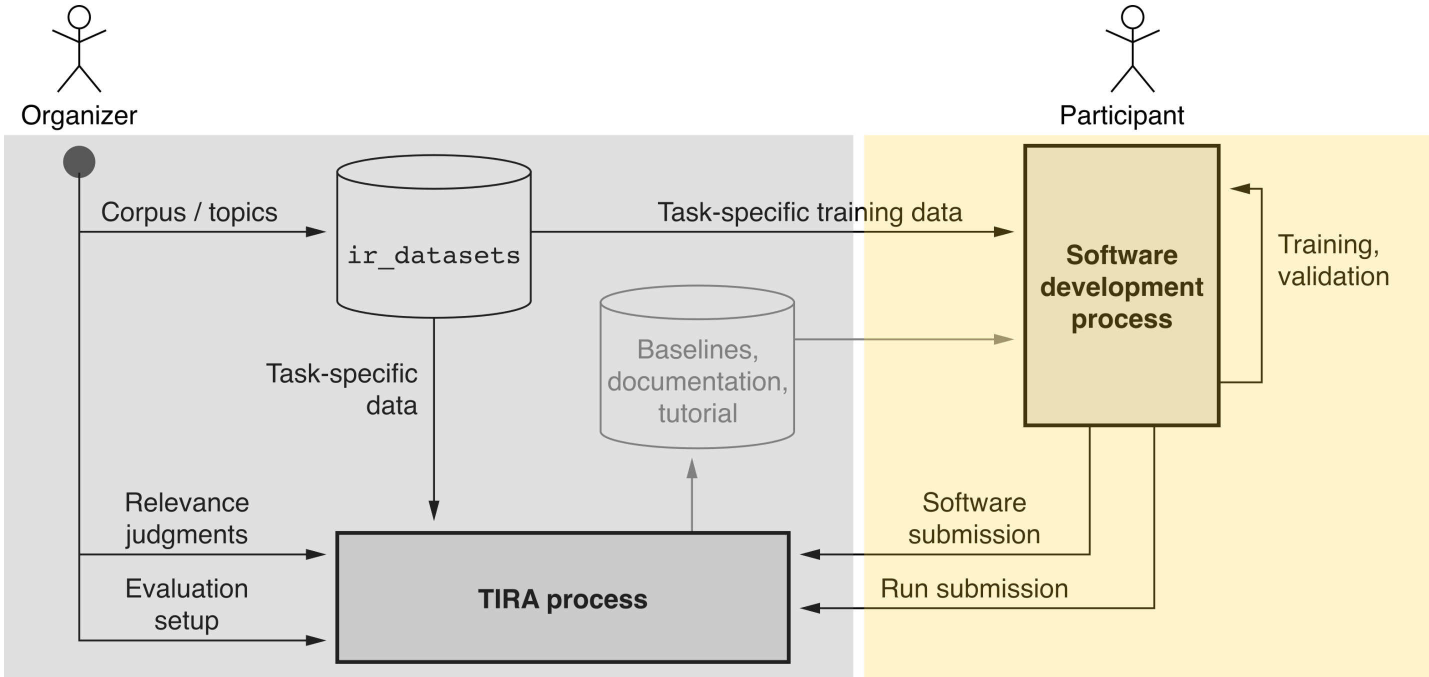
Baselines

- Retrieval, then generation:
  - Provide LLM with query and snippets from a traditional IR system
  - Relies on high context length
- Generation, then retrieval:
  - Use LLM to generate a better query that is piped into a traditional IR system
- Approaches in between
- ChatNoir Chat based on Alpaca and the Clueweb

③

Systems for Reproducible Evaluation Experiments

Adaptation of TIRA for Generative IR Systems



- Participants submit LLM-powered generative IR systems
- Central evaluation on given tasks
- GPU support

Implementing a Reproducible LLM Infrastructure

- Self-hosted LLMs and dynamically changing blackboxes (ChatGPT) are problematic (even with wiretapping)

Approaches should still work
when changing the underlying LLM

Implementing a Reproducible LLM Infrastructure

- Self-hosted LLMs and dynamically changing blackboxes (ChatGPT) are problematic (even with wiretapping)

Approaches should still work
when changing the underlying LLM

- Providing an API infrastructure
- Allows to repeat approaches later by switching to a newer state-of-the-art LLM
- Real time hosting vs. batch processing
- Hosting many different LLMs in parallel is difficult
- Working on a Kubernetes infrastructure for dynamic scaling (scale to zero)

Conclusion

- Adapting traditional IR system evaluation for generative IR systems
- Need to focus on reproducibility when designing evaluation systems

Discussion

- How to solve the problem of having no pre-labeled judgements?
  - Offline evaluation with human annotations? Requires new evaluation for every new version of the underlying LLM
  - LLM-based simulation? Requires an LLM that is *better* than the one underlying the generative IR system

- Where will generative IR systems be extended to, requiring different user models and metrics? Image SERPs, . . .