Towards Query Logs for Privacy Studies: On Deriving Search Queries from Questions

Asia J. Biega[1], Jana Schmidt[2] and **Rishiraj Saha Roy**[2]

[1] Microsoft Research, Canada

[2] Max Planck Institute for Informatics, Germany

42nd European Conference on Information Retrieval (ECIR 2020)

Motivation

- **User-specific search logs** vital for research in privacy and personalization in IR
- Commercial search engines **do not release** such query histories
- **Community question answering (CQA)** platforms a viable alternative for deriving query logs
- Contain information needs that users typically **“translate”** to queries
- **Query formulation** a ubiquitous phenomenon but not well-understood
- Goal is to **understand** query formulation behavior and create a useful **resource** in the process

TL;DR: Visit our page for resource with 7000 (question, query) pairs at https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/mediator-accounts/

The StackExchange Platform

- Contains necessary desiderata for resource creation
  - More than 150 topically diverse sub-forums
  - User IDs shared across these sub-forums: Detailed profiles can be created by stitching
  - Contain valuable signals like accepted answers, related and duplicate questions
- Terms of service allows reuse of data for research purposes
- Details at: https://stackexchange.com/sites
- Curated dumps regularly made available: https://archive.org/details/stackexchange
- But questions in verbose natural language (NL) …

Contributions

- Large-scale crowdsourced study for converting NL questions to search queries
- Conducted on Amazon Mechanical Turk (AMT) with rigorous bias control
- Insights behind query formulation that could drive strategies for conversion at scale
- **Released resource with 7000 question-query pairs spanning 50 domains**
- Available at https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/mediator-accounts/

Setting up the user study (I)

- Filtering sub-forums
  - Excluded those not focused on text (mathematics, programming)
  - Excluded those not focused on English (other languages)
  - Excluded specialized sub-forums that may not be interpretable to Turkers (min. 100 questions)
  - Resulted in 75 sub-forums
- Sampling questions for conversion
  - Proxy for understandability: Presence of accepted answer + five other answers
  - These answers can later be used for creating pseudo-relevant document collections
  - 100 such questions sampled from each sub-forum

Crowdsourcing setup on AMT

- **Task:** Convert NL question to Web search query
- 100 AMT **Masters** with >95% success rate: https://www.mturk.com/help#what_are_masters
- Each Turker converted 50 questions (1 HIT) to capture **user-specific querying behavior**
- **Payment:** $9/hour with **allotted time** 3 hours per HIT (time taken: 1.6 hours on average)
- **Pilot study** validated setup works smoothly
- **Main study** resulted in 5000 question-query pairs
- **Mean query length:** 6.2 words, showing complexity of information needs

Guidelines

- Kept to a minimum to avoid biases towards specific types of queries
- Building search queries aimed at retrieving equivalent information as source question
- Five representative examples provided
- Allowed exact (“use”), modified (“using”) or own words (“utilizing”) for question ➔ query
- Question created by concatenating title, body and sub-forum name (for context)
- Collected demographic information for Turkers

Exact guidelines

When we need to look for some information or answer a question (that is, satisfy an information need), we often turn to a search engine to find the answers. Search systems such Google or Bing return results in response to web search queries that succinctly express our information needs.
In this HIT, we ask you to do the following for 50 examples:
1. read a descriptive **question** that was asked by a person looking for certain information,
2. and then, provide a **search query** you would use if you were to look for the **same information** as the author of the question, but **using a search engine** (such as Google).
To formulate the query, you can either select words from the provided body of text, or use your own words if necessary. Seven solved examples that guide you through representative scenarios are provided for your convenience.

One person is allowed to do only one HIT.
If you have any questions, please contact Rishiraj Saha Roy: rishiraj@mpi-inf.mpg.de

Example

[travel] How do I cancel a Schengen tourist visa? I currently have a valid Schengen tourist visa that expires next year, issued by the French Consulate in Saudi Arabia. I am going to a 60-day study abroad program in Italy and they require a student visa. I went to the Italian Consulate in Boston and they asked me to go to the French Consulate and cancel my visa. However, the French Consulate said that I needed to contact the French Embassy in Saudi to cancel the visa. How can I cancel my tourist visa if I am currently in the United States? My Italian student visa application is due tomorrow morning.

An acceptable query is shown here:
cancel tourist visa french consulate usa
The words have been marked in color only for illustrating where they were chosen from. You only have to provide the query text; no explanation or coloring is necessary.

[travel] How do I cancel a Schengen tourist visa? I currently have a valid Schengen tourist visa that expires next year, issued by the French Consulate in Saudi Arabia. I am going to a 60-day study abroad program in Italy and they require a student visa. I went to the Italian Consulate in Boston and they asked me to go to the French Consulate and cancel my visa. However, the French Consulate said that I needed to contact the French Embassy in Saudi to cancel the visa. How can I cancel my tourist visa if I am currently in the United States? My Italian student visa application is due tomorrow morning.

Analysis

- **Goal:** Discriminate between words that are selected for querying, and those that are not
- **Position**
  - 60% of query words originate from the first 10% of the question
  - Users conceptualize content (core) first, followed by intent (conditions) in questions
  - 17% new words: Challenge for query generative models
- **Part-of-speech (POS)**
  - Content words (nouns, verbs, adj., adv.): 79%, function words (prepositions, conjunctions, …): 21%
  - Reinforcing **content-intent hypothesis**
- **Frequency**
  - Query terms were found to have higher term frequency than non-query terms
- Promising insights for methods aimed at **predicting "query" words**

Control studies

- Two separate control studies to check for spam and annotation “correctness”
- Title position bias
  - **Are Turkers biased by title? Do they look beyond the title?**
  - In 10 HITs (500 questions), title placed at end without informing annotators
  - Large proportion of words chosen from title, even when placed at end – titles do summarize question!
  - Reasonable proportion of query words came from last sentence in general
- Inter-annotator agreement
  - 10 HITs (500 questions) annotated by three workers instead of one
  - 33% Jaccard overlap across annotators – mostly on query “content”
- **Bottomline:** AMT Master workers did task faithfully ☺

Summary and future work

- Released **7000 question-query pairs** for building query generative models
  - 5000 from main study
  - 2000 from control studies
  - Contain worker ID, question topic (sub-forum) and other metadata
- Key insights explaining **query formulation**
- Next steps
  - Use answers as **pseudo-relevant documents**, pool for **corpus**
  - Accepted answers and upvotes / downvotes for **relevance assessments**
- Community question answering forums a rich resource for **simulating query logs**
- Such logs with detailed user profiles can **boost IR research** on privacy and personalization!!