Md Mustafizur Rahman Web: https://www.ischool.utexas.edu/~nahid E-mail: nahid@utexas.edu

Statement of Purpose

University Graduate Continuing Fellowship

Background: With the massive growth of digital information (e.g. the rise of the Web, email, and Twitter, etc.), accurate and efficient search systems such as Google have become critical to the functioning of modern society. While today's corporate search engines already provide search solutions in many areas of commercial interest (e.g., Web search), many important search problems lie beyond commercial interest (e.g., searching oral history). The field of information retrieval (IR) research broadly investigates the science of engineering effective search systems.

Problem Statement: Development of new search algorithms requires an evaluation framework in which A/B testing of new vs. existing algorithms can be reliably performed. While today's search evaluation methodology is reliable, it relies heavily upon people manually annotating the relevance of many search results, which is slow and expensive. Moreover, this practice has become increasingly infeasible as digital collections have grown ever-larger. Consequently, there is an urgent need today for better IR evaluation methods which are both cost-effective and reliable. My doctoral research focuses on developing low-cost yet reliable IR evaluation methods by integrating state-of-the-art artificial intelligence (AI) techniques with traditional human annotation.

Current Work (1 of 2): My first line of research toward this goal was submitted in September 2018 to the Journal of the Association for Information Science and Technology (JASIS&T) [1]. In this work, I seek to reduce the human annotation effort needed to evaluate IR systems by using a sophisticated AI technique, known as *active learning*. Specifically, rather than relying entirely on human annotators to judge search results, I propose an amalgam of human annotation and AI automation. **Figure 1** illustrates the difference between the proposed hybrid model and traditional IR evaluation. This hybrid system works as follows. Firstly, during the initial stages, the system selects a set of search results for human judges to annotate which will be most informative for training the AI model. After training the AI model using these human judgments, we use the AI system to automatically judge the remaining search results. Experimental evaluation over various dataset shows that the AI system achieves 90% accuracy in annotating search results when their human counterparts provide 60% fewer annotations.

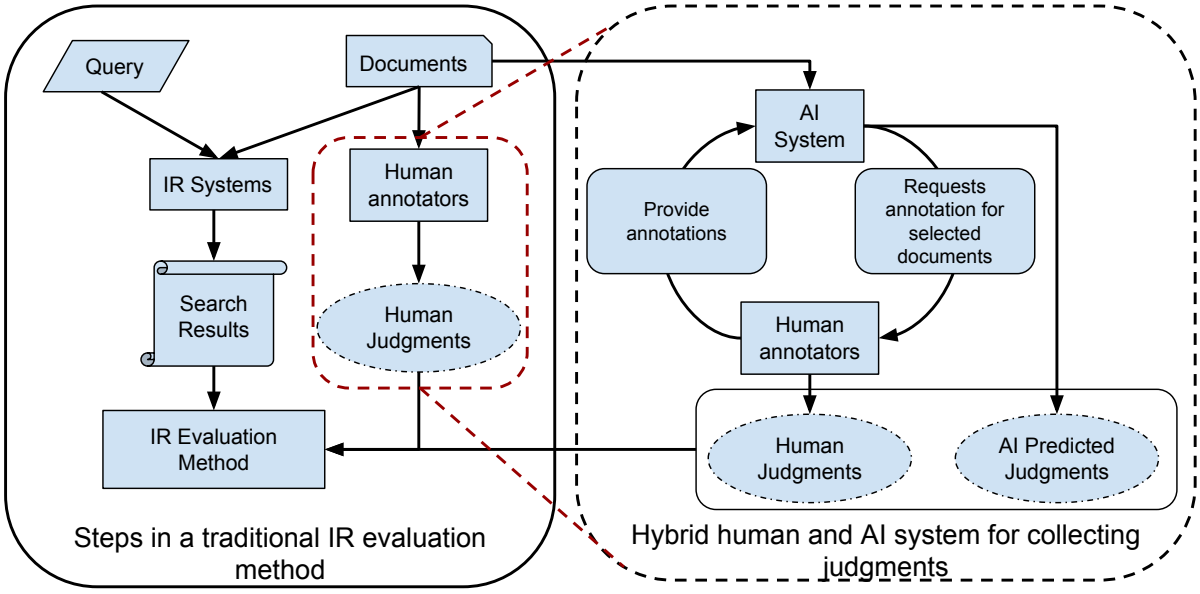

Figure 1: **Left.** Traditional IR evaluation reliant entirely on manual judging of search results. **Right.** My proposed hybrid human and AI system to collect accurate relevant judgments more efficiently.

Current Work (2 of 2): In the aforementioned JASIS&T submission, my method focused on intelligent selection of a minimal set of search results for human annotators to judge for search query. In another line of research, I aim at predicting how human judging effort can be intelligently allocated across different search queries. These two lines of work are complementary: both seek to use human effort as judiciously as possible, maximizing the value of the evaluation dataset produced while simultaneously minimizing the human judging cost required to develop it. Whereas traditional approaches allocate the same human judging effort across different search queries, I utilize an AI technique known as *reinforcement learning* to predict the number of human relevance judgments which should be collected for each search query. Results show that, my proposed approach can find, on average, 90% of the relevant search results using 52% fewer human judgments. This work [2] is currently under review for the World Wide Web (WWW 2019) conference.

Additional Work: While at UT Austin, I have also co-authored several other articles in addition to those described above. One peer-reviewed article [3], published in the Journal of Information Retrieval, extends an earlier technical report [4] in providing a literature review of the state-of-the-art in *deep learning* AI approaches for IR. Another conference paper [5], published at the Association for Computing Machinery (ACM) Computer-Supported Cooperative Work and Social Computing (CSCW) conference, investigates how human judgments of word similarity are impacted by how the textual prompts are presented to the judges. Finally, while not performed while at UT Austin, I note that during my Master of Computer Science at the University of Virginia, I also published a first-authored peer-reviewed article [6] at the 2016 World Wide Web (WWW) conference. This work presents a novel combination of AI modeling techniques (latent Dirichlet allocation and hidden markov models) in order to accurately infer the human sentiment underlying a textual document.

Future Work: As stated at the beginning of the proposal, my doctoral research focuses on developing low-cost IR evaluation methods by integrating AI techniques with human annotation. One future direction I am interested to pursue investigates how people use different queries to search for the same information. Prior work has shown that accuracy of IR systems can vary greatly across different queries, yet current evaluation methods typically only consider a single query for each user's search goal. Consequently, I am investing whether we should allocate more human effort during IR system evaluation in developing a broader set of queries on which to evaluate IR systems (i.e. breadth), rather than current practice of judging many documents for fewer queries (i.e. depth). As another direction of future research, *crowdsourcing* has recently emerged as an alternative, low-cost way to collect many human judgments at a significantly lower cost than required by traditional human labor models. However, crowdsourced human judgments are also often less reliable than those collected via traditional labor practices. Consequently, I would like to investigate whether we can develop an intelligent AI model which can combine resources from both experts and crowd workers and thus reduce costs even further.

Biographical sketch: I am a 3rd year PhD student in the School of Information. Prior to coming to UT Austin, I earned my Master of Computer Science from the University of Virginia. Beyond my research experiences to date at UT Austin, I have sought out and obtained valuable industrial research experiences. During the summer of 2017, at Los Alamos National Laboratory, I developed an AI model to automatically extract clinical information from cancer pathology reports. While at Samsung Research in summer 2018, I worked on improving the question answering module of digital assistants (e.g. Samsung's Bixby, Apple's Siri, etc.) by integrating an AI-based IR system. These experiences have provided me with additional research training and skill development that will enable me to more successfully execute my planned research.

Professional Goals: My long-term career goal is to work as an information retrieval researcher, either as a faculty member at a renowned university or as a research scientist at a respected industry research lab, such as Microsoft Research.

Financial Need: My PhD adviser has supported me as a graduate research assistant (GRA) for the past two and a half years. Although he is actively applying for grants to fund me, currently he has no

grant funding for the next academic year. As a result, the University Graduate Continuing Fellowship would enable me to continue to focus on advancing my research in the next academic year while my adviser seeks grants to fund me for the subsequent years.

References

[1] **M. M. Rahman**, M. Kutlu, T. Elsayed, and M. Lease, "Efficient test collection construction via active learning," *CoRR*, vol. abs/1801.05605, 2018, (under revision for JASIS&T). [Online]. Available: http://arxiv.org/abs/1801.05605

[2] **M. M. Rahman**, M. Kutlu, and M. Lease, "Constructing Test Collections using Multi-armed Bandits and Active Learning," in *Proceedings of the 29th International Conference on World Wide Web (WWW)*, 2019, (under review).

[3] K. D. Onal, Y. Zhang, I. S. Altingovde, **M. M. Rahman**, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. de Rijke, and M. Lease, *Neural Information Retrieval: At The End of The Early Years*. Springer, June 2018, vol. 21, no. 2. [Online]. Available: https://doi.org/10.1007/s10791-017-9321-y

[4] Y. Zhang, **M. M. Rahman**, A. Braylan, B. Dang, H. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, and M. Lease, "Neural information retrieval: A literature review," *CoRR*, vol. abs/1611.06792, 2016. [Online]. Available: http://arxiv.org/abs/1611.06792

[5] M. Bhattacharyya, Y. Suhara, **M. M. Rahman**, and M. Krause, "Possible confounds in word-based semantic similarity test data," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, pp. 147–150. [Online]. Available: http://doi.acm.org/10.1145/3022198.3026357

[6] **M. M. Rahman** and H. Wang, "Hidden Topic Sentiment Model," in *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016, pp. 155–165, (**Acceptance rate: 16%**). [Online]. Available: https://doi.org/10.1145/2872427.2883072