Distribution Agreement

Signature:

______________________________ ______________
Denis Savenkov Date

Question Answering with User Generated Content

By

Denis Savenkov
Doctor of Philosophy

Mathematics and Computer Science

Eugene Agichtein, Ph.D.
Advisor

Jinho D. Choi, Ph.D.
Committee Member

Li Xiong, Ph.D.
Committee Member

Scott Wen-tau Yih, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Question Answering with User Generated Content

By

Denis Savenkov
M.S., Tula State University, 2007

Advisor: Eugene Agichtein, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the Graduate School
of Emory University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Mathematics and Computer Science
2017

Abstract

Question Answering with User Generated Content
By Denis Savenkov

Modern search engines have made dramatic progress in answering many user questions, especially about facts, such as those that might be retrieved or directly inferred from a knowledge base. However, many other more complex factual, opinion or advice questions, are still largely beyond the competence of computer systems. For such information needs users still have to dig into the "10 blue links" of search results and extract relevant information. As conversational agents become more popular, question answering (QA) systems are increasingly expected to handle such complex questions and provide users with helpful and concise information.

In my dissertation I develop new methods to improve the performance of question answering systems for a diverse set of user information needs using various types of user-generated content, such as text documents, community question answering archives, knowledge bases, direct human contributions, and explore the opportunities of conversational settings for information seeking scenarios.

To improve factoid question answering I developed techniques for combining information from unstructured, semi-structured and structured data sources. More specifically, I propose a model for relation extraction from question-answer pairs, the Text2KB system for utilizing textual resources to improve knowledge base question answering, and the EviNets neural network framework for joint reasoning using structured and unstructured data sources. Next, I present a non-factoid question answering system, which effectively combines information obtained from question-answer archives, regular web search, and real-time crowdsourcing contributions. Finally, the dissertation describes the findings and insights of three user studies, conducted to look into how people use dialog for information seeking scenarios and how existing commercial products can be improved, e.g., by responding with certain suggestions or clarifications for hard and ambiguous questions.

Together, these techniques improve the performance of question answering over a variety of different questions a user might have, increasing the power and breadth of QA systems, and suggest promising directions for improving question answering in a conversational scenario.

Question Answering with User Generated Content

By

Denis Savenkov
M.S., Tula State University, 2007

Advisor: Eugene Agichtein, Ph.D.

A dissertation submitted to the Faculty of the Graduate School
of Emory University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Mathematics and Computer Science
2017

Acknowledgments

My very first words of gratitude are addressed towards my advisor Professor Eugene Agichtein for his continuous support of my Ph.D. study, for his limitless patience, motivation, and guidance during these years. My very first interaction with Professor Agichtein happened in 2010 in Petrozavodsk, Russia, where Eugene was teaching a course at the Russian Summer School in Information Retrieval. I was really intrigued by the research done in his lab, and captivated by Eugene's personality traits. Therefore, when I decided to pursue a Ph.D., the application to Emory University was a natural choice. It took me a while to settle on the topic of this dissertation, and Eugene was very supportive throughout the whole time. During our numerous conversations I have learned so many different things from him: how to spot interesting problems, be on top of recent research, structure and write papers, and many others. During my Ph.D. years, Eugene helped me grow not only as a researcher but also as a person. He helped me adopt a healthier lifestyle, revive the love for running, lose 70 lbs (32 kg), and achieve the best physical shape I have ever had. I am extremely fortunate to have Eugene as a mentor and Ph.D. advisor.

Besides my advisor, I would like to thank the rest of my dissertation committee: Prof. Jinho Choi, Prof. Li Xiong, and Dr. Scott Wen-tau Yih, for their time and insightful comments and suggestions, which helped me form and revise this dissertation.

A fraction of this thesis would be impossible without collaborations. The work on relation extraction from question-answer pairs started as an internship project in Google's Knowledge Vault team, where I had a chance to work with Dr. Wei-Lwun Lu as my host, Dr. Evgeniy Gabrilovich, Dr. Jeff Dalton and Dr. Amarnag Subramanya. This experience has taught me a number of lessons, introduced to many great people, and had a big impact on me personally. A large portion of work on conversational question answering was done in collaboration with Prof. Charles L.A. Clarke, Prof. Pavel Braslavsky, and Alexandra Vtyurina. These collaborations helped me get valuable experience and widen my research horizons. I would like to thank all my collaborators and co-authors in various other projects: David Carmel, Nick Craswell, Jeff Dalton, Evgeniy Gabrilovich, Dan Gillick, Qi Guo, Dmitry Lagun, Qiaoling Liu, Wei-Lwun Lu, Yoelle Maarek, Tara McIntosh, Dan Pelleg, Edgar Gonzlez Pellicer, Marta Recasens, Milad Shokouhi, Idan Szpektor, Yu

Wang and Scott Weitzner. While some of these collaborations did not become a part of this thesis, it would be difficult to overstate the impact they had on some of my ideas, skills, and experience.

Special thanks goes to people, who planted and helped me grow the grain of love towards the research, people who were my mentors during the bachelor and masters studies, during the time at the wonderful Yandex School of Data Science and life changing internship and work at Yandex: Sergey Dvoenko, Vadim Mottl, Ilya Muchnik, Dmitry Leschiner and many others.

I would like to thank my fellow labmates, past and present members and visitors of the Emory IRLab: Noah Adler, Liquan Bai, Pavel Braslavsky, David Fink, Qi Guo, Payam Karisani, Alexander Kotov, Dmitry Lagun, Qiaoling Liu, Alexandra Vtyurina, Yu Wang, Zihao Wang and Nikita Zhiltsov. Thank you for enriching my Ph.D. studies with insightful discussions, feedback, support and relaxing chitchats. The years of Ph.D. work would have been unbearable without my friends in Atlanta, who helped to settle in a new country and new city, and supported along the way. I want to thank them for being here for me and for all the fun we have had in the last six years. I hope to carry this friendship over the years.

And finally, I would like to thank people, without whom I could not have achieved anything: my family. I thank my grandmother, who passed away before I started my Ph.D., but who nurtured me and supported all my endeavors since I was 3 years old. I thank my mom, who always believed in me and spiritually contributed to this dissertation as much as I did. I know it has been very hard for her to be that far from her son. And last, but not least, all this would be impossible without my wife, who sacrificed her career to join me in this journey. Words cannot describe my gratitude and love to her. I thank her for all the understanding and support, and for the son, who brought so much joy in our lives, and gave me the energy to work and improve.

This research was funded by the Yahoo! Labs Faculty Research Engagement Program, Google Faculty Research Award, NSF IIS-1018321, DARPA D11AP00269 and by travel support from the ACM Special Interest Group on Information Retrieval and Association for Computational Linguistics.

To my wife Jenny, who is supporting me throughout my career.