Question Answering with Knowledge Base, Web and Beyond

Scott Wen-tau Yih & Hao Ma
Microsoft Research

Introduction

Developing a Question Answering (QA) system to automatically answer natural-language questions has been a long-standing research problem since the dawn of AI, for its clear practical and scientific value. For instance, whether a system can answer questions correctly is a natural way to evaluate a machine's understanding of a domain. Providing succinct and precise answers to informational queries is also the direction pursued by the next generation of search engines that aim to incorporate more "semantics", as well as the basic function in digital assistants like Siri and Cortana.

In this tutorial, we aim to give the audience a coherent overview of the research of question answering. We will first introduce a variety of QA problems proposed by pioneer researchers and briefly describe the early efforts. By contrasting with the current research trend in this domain, the audience can easily comprehend what technical problems remain challenging and what the main breakthroughs and opportunities are during the past half century. For the rest of the tutorial, we select three categories of the QA problems that have recently attracted a great deal of attention in the research community, and will present the tasks with the latest technical survey.

The first two categories regard answering factoid questions, where the main difference of the problem settings is the information source used for extracting answers. *QA with knowledge base* aims to answer natural language questions using real-world facts stored in an existing, large-scale database. The representative approach for this task is to develop a semantic parser (of questions), which will be the main focus. Other approaches like text matching in the embedding space and those driven by information extraction will also be discussed. The other category, *QA with the Web*, targets answering questions using mainly from the facts extracted from general text corpora derived from the Web. In addition to the common components and techniques used in this setting, including passage retrieval, entity recognition and question analysis, we will also introduce latest work on how to leverage and incorporate additional structured and semi-structured data to improve the performance. The third category of the QA problems that we will highlight is the non-factoid questions. Due to its broad coverage, we will briefly cover three exemplary topics: story comprehension, reasoning questions and paragraph QA. The tutorial will conclude by summarizing a whole area of exciting and dynamic research that is worthy of more detailed investigation for many years to come.

Outline

Part I. Overview of Question Answering Research

- Overview of early Question Answering research
    - Natural language understanding problems proposed at the dawn of AI

Proceedings of NAACL-HLT 2016, pages 8–10,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

    - Early representative QA systems
    - Key developments and milestones
- Current Question Answering research trend
    - Categories of QA problems and settings studied recently
    - Data sources, technical problems and solutions
    - Main challenges and opportunities
- Demos of some existing QA systems

Part II. Question Answering with Knowledge Base

- Introduction to modern large-scale knowledge base
- Task setting and benchmark datasets
- State-of-the-art approaches
    - ▪ Semantic parsing (of questions)
    - ▪ Matching questions and answers in embedding space
    - ▪ Information extraction and text matching

Part III. Question Answering with the Web

- Problem setting and the general system architecture
- Essential natural language analysis: entity and answer type
- Leveraging additional information sources
    - Usage data (e.g., search query logs or browsing logs)
    - Knowledge bases
    - Semi-structured data (e.g., Web tables)

Part IV. Non-Factoid Question Answering

- Story comprehension (e.g., MC-Test)
- Reasoning questions (e.g., bAbI dataset & task)
- Paragraph QA (e.g., quiz bowl competition)

Part V. Conclusion

Instructor bios

Scott Wen-tau Yih is a Senior Researcher at Microsoft Research Redmond. His research interests include natural language processing, machine learning and information retrieval. Yih received his Ph.D. in computer science at the University of Illinois at Urbana-Champaign. His work on joint inference using integer linear programming (ILP) [Roth & Yih, 2004] helped the UIUC team win the CoNLL-05 shared task on semantic role labeling, and the approach has been widely adopted in the NLP community since then. After joining MSR in 2005, he has worked on email spam filtering, keyword extraction and search & ad relevance. His recent work focuses on continuous semantic representations using neural networks and matrix/tensor decomposition methods, with applications in lexical semantics, knowledge base embedding and question answering. Yih received the best paper award from CoNLL-2011, an outstanding paper award from ACL-2015 and has served as area chairs (HLT-NAACL-12, ACL-14, EMNLP-16), program co-chairs (CEAS-09, CoNLL-14) and action editor (Transactions of ACL) in recent years.

Hao Ma is a Researcher at Microsoft Research, Redmond, WA, USA. He obtained his Ph.D. in Computer Science at The Chinese University of Hong Kong. His research interests include Information Retrieval, Natural Language Processing, Machine Learning, Recommender Systems and Social Network Analysis. Most recently, Dr. Ma has been working on entity related research problems and applications. He designed the core learning algorithms that powered both Bing's and Microsoft's entity experience, including question answering, entity recommendation, attributes ranking, interpretation, exploration, carousel ranking, etc. He has published more than 40 research papers in prestigious conferences and journals, including WWW, SIGIR, WSDM, AAAI, TOIS, TKDE, TMM, TIST, etc. Some of his research work has been widely reported by popular news media, like MIT Technology Review, Search Engine Land, etc. Dr. Ma is also in the winning team that won the Microposts Entity Linking Challenge in WWW 2014.