Question Answering with Knowledge Base, Web and Beyond

Wen-tau Yih
Microsoft Research
One Microsoft Way
Redmond, WA, USA
scottyih@microsoft.com

Hao Ma
Microsoft Research
One Microsoft Way
Redmond, WA, USA
haoma@microsoft.com

ABSTRACT

In this tutorial, we give the audience a coherent overview of the research of question answering (QA). We 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 present the tasks with the latest technical survey. We conclude the tutorial by discussing the new opportunities and future directions of QA research.

1. 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. The tutorial starts from the historical background, but puts more emphasis on recently emerging approaches, such as leveraging structured and semi-structured information sources, in addition to a large-scale corpus. We select three categories of the QA problems that have recently attracted a great deal of attention in the research community. 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. Examples of the most recent KB-based QA systems include [1, 2, 4, 18, 22, 25, 24]. The representative approach for this task is to develop a semantic parser (of questions), which is the main focus. Other approaches like text matching in the embedding space and those driven by information extraction are also discussed. The other category, QA with the Web [3, 7, 11, 19, 17], 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 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 highlight is the non-factoid questions. Due to its broad coverage, we briefly cover three exemplary topics: story comprehension [15], reasoning questions [21, 12] and paragraph QA [10].

This tutorial aims to present the current state of research in the emerging question answering field. After the completion of the tutorial, the expected learning outcomes are:

1. The audience will acquire the basic understanding and overview of the emerging topics of question answering.
2. The audience will learn and understand some of the current research work, as well as industry practices using IR, NLP and ML techniques in question answering. The audience will also be able to incorporate some aspects of what has been learned into their own work.
3. The audience will be able to identify a few research directions that could have big impact in the near future.

2. RELEVANCE

The topic of Question Answering is timely for the IR community. QA has again emerged in the past few years, as the trend in the web development and growth is clearly to make search results more semantic. Many research papers published in SIGIR [11, 19], WWW [16, 17, 20] and ACL [2, 3, 5, 23, 24] studied QA-related problems in-depth recently.

To our knowledge, no recent Question Answering tutorial has been delivered in related venues, such as SIGIR, WWW and ACL. Earlier tutorials on corpus-based QA methods were presented by Sanda Harabagiu and Dan Moldovan in NAACL-2001 [9] and by Jimmy Lin and Boris Katz in EACL-2003 [13]. Tutorials dedicated to the IBM Watson system were given more recently [8, 6], although they did not intend to cover more general topics in QA research. In a related WWW-2015 tutorial titled "An Introduction to Entity Recommendation and Understanding" [14], the presenters Hao

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914804

Ma and Yan Ke briefly introduced Question Answering in the context of Entity Recommendation and Understanding. In comparison, this tutorial covers more in-depths introduction to QA-related problems and solutions.

3. FORMAT AND DETAILED SCHEDULE

The main content of this tutorial consists of the historical background of the question answering research, detailed description of open-domain factoid question answering approaches using structured and unstructured information sources, and introduction to recent, non-factoid question answering tasks. The outline is listed below.

Part I. Overview of Question Answering Research.

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

- Summary of the tutorial
- Directions of future research

4. REFERENCES

[1] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544, 2013.

[2] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Proceedings of ACL*, 2014.

[3] E. Brill, S. Dumais, and M. Banko. An analysis of the askmsr question-answering system. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 257–264. Association for Computational Linguistics, 2002.

[4] A. Fader, L. Zettlemoyer, and O. Etzioni. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014.

[5] A. Fader, L. S. Zettlemoyer, and O. Etzioni. Paraphrase-driven learning for open question answering. In *ACL (1)*, pages 1608–1618, 2013.

[6] J. Fan and K. Barker. Natural language processing in watson. In *AAAI-2015: Tutorials*, 2015.

[7] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[8] A. M. Gliozzo, A. Kalyanpur, and J. Fan. Natural language processing in watson. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, page 4. Association for Computational Linguistics, 2012.

[9] S. Harabagiu and D. Moldovan. Open-domain textual question answering. In *Proceedings of the 2001 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Association for Computational Linguistics, 2001.

[10] M. Iyyer, J. L. Boyd-Graber, L. M. B. Claudino, R. Socher, and H. Daumé III. A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644, 2014.

[11] J. Ko, E. Nyberg, and L. Si. A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–350. ACM, 2007.

[12] M. Lee, X. He, W. tau Yih, J. Gao, L. Deng, and P. Smolensky. Reasoning in vector space: An exploratory study of question answering. In *Proceedings of the International Conference on Learning Representations (ICLR) 2016*, 2016.

[13] J. Lin and B. Katz. Question answering techniques for the world wide web. *EACL-2003 Tutorial*, 2003.

[14] H. Ma and Y. Ke. An introduction to entity recommendation and understanding. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 1521–1522. International World Wide Web Conferences Steering Committee, 2015.

[15] M. Richardson, C. J. Burges, and E. Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, volume 1, page 2, 2013.

[16] H. Sun, H. Ma, X. He, W.-t. Yih, Y. Su, and X. Yan. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[17] H. Sun, H. Ma, W.-t. Yih, C.-T. Tsai, J. Liu, and M.-W. Chang. Open domain question answering via semantic enrichment. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1045–1055, 2015.

[18] C. Unger, L. Bühmann, J. Lehmann, A.-C. Ngonga Ngomo, D. Gerber, and P. Cimiano. Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648. ACM, 2012.

[19] E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM, 2000.

[20] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526, 2014.

[21] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *Proceedings of the International Conference on Learning Representations (ICLR) 2016*, 2016.

[22] M. Yahya, K. Berberich, S. Elbassuoni, M. Ramanath, V. Tresp, and G. Weikum. Natural language questions for the web of data. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 379–390. Association for Computational Linguistics, 2012.

[23] X. Yao and B. Van Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, 2014.

[24] W.-t. Yih, M.-W. Chang, X. He, and J. Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1321–1331, Beijing, China, July 2015. Association for Computational Linguistics.

[25] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. Zhao. Natural language question answering over rdf: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 313–324. ACM, 2014.