Full text available at: http://dx.doi.org/10.1561/1500000001

Open-Domain Question–Answering

Open-Domain Question–Answering

John Prager

IBM T.J. Watson Research Center
Yorktown Heights
NY 10598
USA
jprager@us.ibm.com

the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is J. Prager, Open-Domain Question–Answering, Foundation and Trends® in Information Retrieval, vol 1, no 2, pp 91–231, 2006

ISBN: 978-1-60198-052-6

Foundations and Trends® in Information Retrieval

Volume 1 Issue 2, 2006

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians
Foundations and Trends® in Information Retrieval, 2006, Volume 1, 4 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 1, No. 2 (2006) 91–231

DOI: 10.1561/1500000001

Open-Domain Question–Answering

John Prager

IBM T.J. Watson Research Center, 1S-D56, P.O. Box 704, Yorktown Heights, NY 10598, USA, jprager@us.ibm.com

Abstract

The top-performing Question–Answering (QA) systems have been of two types: consistent, solid, well-established and multi-faceted systems that do well year after year, and ones that come out of nowhere employing totally innovative approaches and which out-perform almost everybody else. This article examines both types of system in depth. We establish what a "typical" QA-system looks like, and cover the commonly used approaches by the component modules. Understanding this will enable any proficient system developer to build his own QA-system. Fortunately there are many components available for free from their developers to make this a reasonable expectation for a graduate-level project. We also look at particular systems that have performed well and which employ interesting and innovative approaches.

Contents

1

Introduction

Question–Answering (QA) is a research activity which is difficult to define precisely, but most practitioners know what it is when they see it. Loosely speaking, it is the field of study concerning the development of automatic systems to generate answers to questions in natural language. The source of the answers and the manner of generation are left open, as are the kinds of questions. However, as a first approximation, the field is currently mostly concerned with answering factual questions (questions about agreed, or at least authoritatively reported facts) by consulting one or more corpora of textual material.

This is not to say that such questions are exclusively about simple properties of objects and events (the height of Mt Everest, the birthdate of Mozart, and so on). The field is also interested in definitions (finding important unspecified characteristics of an entity); relationships (how entities are interrelated); and even opinions (how people or organizations have reacted to events). What is common between these is that QA systems are currently extractive: They just report information that is found in external resources such as newswire, without any attempt to prove the authors correct, and also without any attempt to construct answers that are only implicit in the sources.

The kind of QA that will be the subject of this article for the most part will be that which is the subject of the annual TREC (Text Retrieval Conference) evaluation at NIST [76] beginning in 1999. The majority of the QA systems that have been developed, both in academia and industrial research labs, have been at least partly for participation at TREC, and the majority of technical papers on the subject have used TREC corpora, question sets and metrics for evaluation, so such a focus is only natural. However, we will also address aspects of QA that TREC has avoided so far, and we will examine some of the deficiencies of the TREC-style approach.

QA draws upon and informs many of the subfields of Information Retrieval (IR) and Natural Language Processing (NLP), but is quite different from them in certain ways. QA is a very practical activity — more an engineering field than a science — and as such, at least today, is more a collection of tools and techniques than formulas and theorems. This is very understandable when one considers that at its heart, QA is concerned with matching a natural language question with a snippet of text (or in general, several snippets), an algorithmic solution to which could be said to be NLP-complete.[1]

QA is heavily reliant on processes such as named entity recognition (NER), parsing, search, indexing, classification and various algorithms from machine learning, but as we will see it seems to be surprisingly insensitive to the particular choices made. In the author's experience, choosing to use a state-of-the-art component over a less advanced version does not usually make much difference in the QA-system's overall performance. What makes a difference is how the components are organized relative to each other; in other words, it is what the system is trying to do that is typically more important than how it does it. This leads to the fascinating situation where contributions usually come from the introduction of brand-new approaches, rather than the fine-tuning of parameters in, say, ranking algorithms. The top-performing systems in TREC have been of two types: consistent, solid, well-established, and multi-faceted systems that do well year after year, and ones that come

[1] A play on the notion of NP-completeness from the field of computational complexity, and AI-completeness, the less formal but still widely held belief that solution to any hard Artificial Intelligence (the parent field of NLP) problem leads to the solution of any other.

out of nowhere employing totally innovative approaches and which outperform almost everybody else.

This article will examine both types of system in depth. We will establish what a "typical" QA-system looks like, and cover the commonly used approaches by the component modules. Understanding this will enable any proficient system developer to build his own QA-system. Fortunately, there are many components available for free from their developers to make this a reasonable expectation for a graduate-level project. We will also look at particular systems that have performed well and which employ interesting and innovative approaches, but we will not examine every single system that has acquitted itself well. We will not cover commercial systems that have not been forthcoming about their internal workings.

1.1 A Brief History of QA

The field of QA, as it is currently conceived, was inaugurated in 1999 when NIST introduced a Question–Answering track for TREC-8. However, it was not for this that the first question–answering systems were developed. For those, we need to go back to the 60s and 70s and the heyday of Artificial Intelligence facilities such as MIT's AI Lab. In those days and in those locations almost all programming was done in LISP and PROLOG and derived languages such as Planner and Micro-Planner. These were the test-beds of pioneering AI systems, many of which could now with hindsight be called QA systems, although they were not at the time (see, e.g., SHRDLU [80]).

For the most part, these systems were natural-language interfaces to databases. A question, problem or action to be taken was input in English. This was parsed into a semantic form — a semantic representation of the "meaning" of the information need. Then either directly or through a theorem-proving or other inferencing system, goals were generated which could be directly translated into database queries or robot commands. The repertoires, both in terms of actions taken or inputs understood, were severely limited, and were either never used in a practical system, or were only usable for the very narrow application for which they were designed. Such systems included LIFER/LADDER [23], LUNAR [81], and CHAT-80 [79].

What these systems had in common was that they were toy systems. They were brittle, and did not scale. They used very complex approaches (inferencing, subgoals, etc.) and did not degrade gracefully [41]. They suffered from lack of general-purpose community resources, which made them expensive to develop or extend. What is ultimately damning is that there was no easily identifiable line of evolution from those systems to the present day; they died out like dinosaurs.

Possibly the first system that can be recognized as what some might call a modern QA system, in the sense that it was open-domain and used unrestricted free text, was the MURAX system [30]. It processed natural-language questions seeking noun-phrase answers from an on-line encyclopedia; it used shallow linguistic processing and IR, but did not use inferencing or other knowledge-based techniques.

The next milestone came shortly after the creation of the World Wide Web: MIT's Start system [27, 28] was the first Web Question–Answering system. This work has progressed to the present day, and is still available.[2] Ask Jeeves[3] (now Ask.com), founded in 1996, was maybe the first widely-known Web QA system, although since it returns documents, not answers, one can debate if it is a true QA system.[4] Since that time, other QA systems have come online, both academic and commercial; these include Brainboost[5] and AnswerBus,[6] both of which return single sentences. There is a strong argument that even if a number or a noun-phrase, say, is the technically correct answer to a question, a sentence attesting to the subject fact is even better. Especially when typical system's accuracy is far from 100%, as much transparency is desired as possible.

In 1992 NIST, the U.S. National Institute of Standards and Technology, inaugurated the annual Text Retrieval Conference, commonly

[2] http://www.start.csail.mit.edu.
[3] http://www.ask.com.
[4] Although if the answer is embedded in the document abstract presented in the hit-list, the end-user typically would not care.
[5] http://www.brainboost.com.
[6] http://www.answerbus.com/index.shtml.

called TREC. Every year, TREC consists of a number of tracks (the exact composition usually changes a little from year to year), each one concerned with a different aspect of IR. In each track, one or more evaluations are run in which teams from around the world participate. TREC is generally now considered a hugely important factor in IR research, since it provides relevance judgments[7] and allows researchers to compare methodologies and algorithms on a common testbed. Many more details about TREC can be found on its official website[8] or in a recently-published book from NIST [76].

In 1999, NIST added a QA track to TREC. There was a feeling that the Question–Answering problem could benefit from bringing together the NLP and IR communities. NLP techniques were, and still are, much more precise than IR techniques, but considerably more computationally expensive; NLP was typically used in closed-world domains, IR typically in open-domain. Thus using the power of IR to search many megabytes or gigabytes of text in short times, together with the refinement of NLP techniques to pinpoint an answer was expected to be a worthwhile technological challenge. The track has continued to this day, although it has changed in ways discussed later.

To emphasize the world-wide interest in QA that has arisen in recent years, we will mention here some other venues for QA. In 2001, NTCIR,[9] a series of evaluation workshops to promote research in IR and NLP activities in Asian languages, introduced a Question–Answering task. In 2003, the Cross-Language Evaluation Forum (CLEF),[10] an European TREC-like context for cross-language IR, inaugurated a multiple-language Question–Answering track. Both of these are ongoing. Recent workshops include: Open-Domain QA (ACL 2001), QA: Strategy and Resources (LREC 2002), Workshop on Multilingual Summarization and QA (COLING 2002), Information Retrieval for QA (SIGIR 2004), Pragmatics of QA (HTL/NAACL 2004), QA in Restricted Domains (ACL 2004), QA in Restricted Domains (AAAI 2005), Inference for Textual QA (AAAI 2005), Multilingual QA (EACL

[7] In some cases provided by NIST assessors, in others by the research community.
[8] http://trec.nist.gov/.
[9] http://research.nii.ac.jp/ntcir/outline/prop-en.html.
[10] http://www.clef-campaign.org/.

2006), Interactive QA (HLT/NAACL 2006) and Task-Focused Summarization and QA (COLING/ACL 2006).

1.1.1 TREC Minutiae

For those interested, Table 3.2 in Section 3.6 lists the teams/systems that have placed in the top 10 in the main QA task since the TREC8 QA in 1999. However, in the remainder of this article we will not for the most part be reporting performance scores of teams since over time these wane in significance, and besides they can be discovered in the teams' own writings and in the annual TREC proceedings. We will report, where known and of interest, how different components contribute to teams' overall scores.

As discussed in [53], the difficulty of questions cannot be assessed independently of knowing the corpus or other resource in which the answers must be found. Up to the writing of this article, TREC has favored newswire text, which has the characteristics of being written by educated English-speaking adults and of being edited, so there is a minimum of typological errors (as compared with mail and Web documents, and especially compared with OCR (optical character recognition) or ASR (automatic speech recognition) documents). Details of the TREC datasets are given in Section 3.6.1.

1.1.2 AQUAINT

In 2001, the U.S. Government, through a small agency called ARDA (Advanced Research Development Activity) began a three-phase multi-year activity called AQUAINT (Advanced Question–Answering for INTelligence). In each phase, a couple of dozen (approximately) U.S. teams, from academia primarily but also industry, were funded to perform research and development with the ultimate goal of producing QA systems that could be used effectively by intelligence analysts in their daily work. One of the goals of the program was to push research away from *factoid* QA (see Section 2.1.1) into questions about reasons, plans, motivations, intentions, and other less tangible quantities. These are very difficult objectives and it remains to be seen to what extent they can be achieved in the near future. One direct consequence of this thrust, though, has been to support and influence the QA-track in TREC.

In particular, the AQUAINT program organized several pilot studies in different dimensions of QA. The Definition Pilot explored a different approach to definition questions — finding a comprehensive set of descriptive text fragments (called *nuggets*) rather than a single phrase as required for factoid questions. When in 2003 TREC started using *question series* — groups of questions about a single *target* — the last question in every series was "other," meaning "return everything important that has not been asked about yet"; this was a direct outgrowth of the Definition pilot.

The Relationship Pilot, where a relationship was defined as one of eight broad ways in which one entity could influence another (organizational, familial, financial etc.) became a subtask of [75, 76]. There have also been Opinion and Knowledge-based question Pilots, which have prompted research reported elsewhere in the literature. More information about these pilots can be found on the NIST web site.[11]

1.2 Article Plan

This article is designed to give readers a good background in the field of Question–Answering. The goal in writing it has been to cover the basic principles of QA along with a selection of systems that have exhibited interesting and significant techniques, so it serves more as a tutorial than as an exhaustive survey of the field. We will not cover (except for occasional mentions in passing) Opinion or Relationship Questions, Interactive QA or Cross-Language QA. Further reading can be found in two recent books [42, 72], as well as the proceedings of the QA tracks of TREC. Finally, we should mention the review article by Hirschman and Gaizauskas [25], which summarizes the state of Question–Answering as of 2001. As of this writing, there is no web-site or umbrella publication from AQUAINT.

The rest of this article is organized as follows. In Chapter 2, we provide a general overview of the theory and practice of

[11] http://trec.nist.gov/data/qa/add_qaresources.html.

Question–Answering (mostly practice). We look at the typical architecture of a QA system, the typical components that comprise it, the technical issues they tackle and some of the most common and successful techniques used to address these problems. In Chapter 3, we look at the different ways QA systems are evaluated. In Chapter 4, we look at some of the specific approaches that have been used by well-performing systems. We will note that three themes seem to permeate these approaches: testing of identity, use of analogy, and detection of redundancy. Because these are very high-level concepts, each of which can be achieved in a number of different ways, it should be of no surprise that different methodologies, namely linguistic, statistical, and knowledge-based, are all found in QA systems. In Chapter 5, we step back and look at the more abstract concepts of User Modeling and Question Complexity; the issues here have not to date been tackled seriously by the community, but it is asserted here that they are of significant importance, and dealing with them will be necessary for future success. We conclude in Chapter 6 with some comments about challenges for QA.

References

[1] A. T. Arampatzis, T. Tsoris, C. H. A. Koster, and T. P. van der Weide, "Phrase-based information retrieval," *Information Processing and Management*, vol. 34, no. 6, pp. 693–707, 1998.

[2] A. L. Berger, V. D. Pietra, and S. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, 1996.

[3] D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," *Machine Learning*, vol. 34, no. 1–3, pp. 211–231, 1999.

[4] S. Blair-Goldensohn, K. R. McKeown, and A. H. Schlaikjer, "Answering definitional questions: A hybrid approach," in *New Directions in Question Answering*, (M. Maybury, ed.), pp. 47–57, AAAI press, 2004.

[5] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng, "Data-intensive question–answering," in *Proceedings of the 10th Text Retrieval Conference (TREC2001)*, NIST, Gaithersburg, MD, 2002.

[6] C. Buckley, M. Mitra, J. A. Walz, and C. Cardie, "SMART high precision: TREC 7," in *Proceedings of the 7th Text Retrieval Conference (TREC7)*, pp. 230–243, Gaithersburg, MD, 1998.

[7] J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C.-Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Srihari, T. Strzalkowski, E. Voorhees, and R. Weishedel, "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) ARDA QA Roadmap (http://www-nlpir.nist.gov/projects/duc/papers/qa, Roadmap-paper_v2.doc)," 2001.

[8] R. Byrd and Y. Ravin, "Identifying and extracting relations in text," in *Proceedings of 4th International Conference on Applications of Natural Language and Information Systems (NLDB 99)*, Klagenfurt, Austria, 1999.

[9] J. G. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR98)*, Melbourne, Australia, August 1998.

[10] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer, "Searching XML documents via XML fragments," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[11] J. Chu-Carroll, J. Prager, C. Welty, K. Czuba, and D. Ferrucci, "A multi-strategy and multi-source approach to question answering," in *Proceedings of the 11th Text Retrieval Conference (TREC2002)*, Gaithersburg, MD, 2003.

[12] J. Chu-Carroll, J. M. Prager, K. Czuba, D. Ferrucci, and P. Duboue, "Semantic search via XML fragments: A high precision approach to IR," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR06)*, Seattle, WA, 2006.

[13] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam, "Exploiting redundancy in question answering," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR01)*, New Orleans, September 2001.

[14] T. Clifton, A. Colquhoun, and W. Teahan, "Bangor at TREC 2003: Q&A and genomics tracks," in *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, Gaithersburg, MD, 2004.

[15] H. Cui, M.-Y. Kan, and T.-S. Chua, "Unsupervised learning of soft patterns for generating definitions from online news," in *Proceedings of the Thirteenth World Wide Web conference (WWW 2004)*, pp. 90–99, New York, May 17–22 2004.

[16] H. Cui, K. Li, R. Sun, T.-S. Chua, and M.-Y. Kan, "National University of Singapore at the TREC-13 question answering main task," in *Proceedings of the 13th Text Retrieval Conference (TREC2004)*, Gaithersburg, MD, 2005.

[17] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[18] A. Echihabi and D. Marcu, "A noisy-channel approach to question answering," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pp. 16–23, 2003.

[19] J. Gao, J.-Y. Nie, G. Wu, and G. Cao, "Dependence language model for information retrieval," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR04)*, Sheffield, UK, July 25–29, 2004.

[20] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley, "Answer mining by combining extraction techniques with abductive reasoning," in *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, Gaithersburg, MD, 2004.

[21] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu, "FALCON: Boosting knowledge for answer engines," in *Proceedings of the 9th Text Retrieval Conference (TREC9)*, pp. 479–488, NIST, Gaithersburg MD, 2001.

[22] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu, "The role of lexico-semantic feedback in open-domain textual question-answering," in *Proceedings of the Association for Computational Linguistics*, pp. 274–281, July 2001.

[23] G. G. Hendrix, E. D. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data," *Very Large Data Bases (VLDB)*, p. 292, 1977.

[24] U. Hermjakob, A. Echihabi, and D. Marcu, "Natural language based reformulation resource and web exploitation for question answering," in *Proceedings of the 11th Text Retrieval Conference (TREC2002)*, Gaithersburg, MD, 2003.

[25] L. Hirschman and R. Gaizauskas, "Natural language question answering: The view from here," *Natural Language Engineering*, vol. 7, no. 4, pp. 275–300, 2001.

[26] I. Ittycheriah, M. Franz, and S. Roukos, "IBM's statistical question-answering system — TREC-10," in *Proceedings of the 10th Text Retrieval Conference (TREC10)*, pp. 258–264, NIST, Gaithersburg, MD, 2001.

[27] B. Katz, "Annotating the world wide web using natural language," in *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet*, 1997.

[28] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. J. McFarland, and B. Temelkuran, "Omnibase: Uniform access to heterogeneous data for question answering," in *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB)*, 2002.

[29] B. Katz and J. Lin, "Selectively using relations to improve precision in question answering," in *Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering*, April 2003.

[30] J. Kupiec, "Murax: A robust linguistic approach for question answering using an on-line encyclopedia," in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR93)*, pp. 181–190, 1993.

[31] C. Kwok, O. Etzioni, and D. S. Weld, "Scaling question answering to the web," in *Proceedings of the 10th World Wide Web Conference (WWW 2001), Hong Kong*, pp. 150–161, 2001.

[32] D. B. Lenat, "Cyc: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, 1995.

[33] X. Li and D. Roth, "Learning question classifiers: The role of semantic information," *Journal of Natural Language Engineering*, vol. 12, no. 3, pp. 229–249, 2006.

[34] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of the Human Language Technology Conference (HLT/NAACL)*, 2003.

[35] J. Lin, "An exploration of the principles underlying redundancy-based factoid question answering," *ACM Transactions on Information Systems*, in press, 2007.

[36] J. Lin and D. Demner-Fushman, “Automatically evaluating answers to definition questions,” in *Proceedings of the Human Language Technology Conference (HLT/EMNLP2005)*, Vancouver, Canada, October 2005.

[37] J. Lin and D. Demner-Fushman, “Will pyramids built of nuggets topple over?,” in *Proceedings the Human Language Technology Conference (HLT-NAACL2006)*, New York, NY, June 2006.

[38] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger, “What makes a good answer? The role of context in question–answering,” in *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, Zurich, Switzerland, 2003.

[39] Linguistic Data Consortium (LDC), “ACE Phase 2: Information for LDC Annotators,” http://www.ldc.upenn.edu/Projects/ACE2, 2002.

[40] B. Magnini, M. Negri, R. Prevete, and H. Tanev, “Is it the right answer? exploiting web redundancy for answer validation,” *Association for Computational Linguistics 40th Anniversary Meeting (ACL-02)*, University of Pennsylvania, Philadelphia, pp. 425–432, July 7–12, 2002.

[41] D. Marr, “Early processing of visual information, philosophic,” *Transactions of the Royal Soc IJT B*, vol. 275, pp. 1377—1388, 1976.

[42] M. Maybury, ed., *New Directions in Question Answering*, AAAI press, 2004.

[43] D. Metzler and W. B. Croft, “A Markov Random field model for term dependencies,” in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pp. 472–479, 2005.

[44] R. Mihalcea and D. Moldovan, “A method for word sense disambiguation of unrestricted text,” in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 152–158, College Park, MD, 1999.

[45] G. Miller, “WordNet: A lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[46] D. Moldovan, C. Clark, S. Harabagiu, and S. Maiorano, “COGEX: A logic prover for question–answering,” in *Proceedings of the Human Language Technology Conference (HLT-NAACL03)*, pp. 87–93, 2003.

[47] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, and V. Rus, “The structure and performance of an open-domain question answering system,” in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL00)*, pp. 563–570, 2000.

[48] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus, “LASSO: A tool for surfing the answer net,” in *Proceedings of the 8th Text Retrieval Conference (TREC8)*, pp. 175–183, Gaithersburg, MD, 1999.

[49] D. I. Moldovan and V. Rus, “Logic form transformation of wordnet and its applicability to question answering,” in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL01)*, Toulouse, France, July 2001.

[50] A. Nenkova and R. Passonneau, “Evaluating content selection in summarization: The pyramid method,” in *Proceedings of the Human Language Technology Conference (NAACL-HLT)*, 2004.

[51] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02)*, 2002.
[52] M. Pasca and S. Harabagiu, "High performance question/answering," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2001)*, pp. 366–374, New Orleans LA, September 2001.
[53] J. M. Prager, "A curriculum-based approach to a QA roadmap," in *LREC 2002 Workshop on Question Answering: Strategy and Resources*, Las Palmas, May 2002.
[54] J. M. Prager, E. W. Brown, A. Coden, and R. Radev, "Question answering by predictive annotation," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR00)*, pp. 184–191, Athens, Greece, 2000.
[55] J. M. Prager, J. Chu-Carroll, E. W. Brown, and K. Czuba, "Question answering by predictive annotation," in *Advances in Open-Domain Question-Answering*, (T. Strzalkowski and S. Harabagiu, eds.), pp. 307–347, Springer, 2006.
[56] J. M. Prager, J. Chu-Carroll, and K. Czuba, "Statistical answer-type identification in open-domain question–answering," in *Proceedings of the Human Language Technology Conference (HLT 2002)*, San Diego, CA, March 2002.
[57] J. M. Prager, J. Chu-Carroll, and K. Czuba, "A multi-agent approach to using redundancy and reinforcement in question–answering," in *New Directions in Question Answering*, (M. Maybury, ed.), pp. 237–252, AAAI press, 2004.
[58] J. M. Prager, J. Chu-Carroll, and K. Czuba, "Question–answering using constraint satisfaction: QA-by-dossier-with-constraints," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL04)*, pp. 575–582, Barcelona, Spain, 2004.
[59] J. M. Prager, J. Chu-Carroll, K. Czuba, C. Welty, A. Ittycheriah, and R. Mahindru, "IBM's PIQUANT in TREC2003," in *Proceedings of the 12h Text Retrieval Conference (TREC2003)*, Gaithersburg, MD, 2004.
[60] J. M. Prager, P. Duboue, and J. Chu-Carroll, "Improving QA accuracy by question inversion," in *COLING-ACL 2006*, pp. 1073–1080, Sydney, Australia, 2006.
[61] J. M. Prager, S. K. K. Luger, and J. Chu-Carroll, "Type nanotheories: A framework for type comparison," *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007)*, to appear.
[62] J. M. Prager, D. R. Radev, and K. Czuba, "Answering what-is questions by virtual annotation," in *Proceedings of Human Language Technologies Conference (HLT)*, San Diego, CA, March 2001.
[63] V. Punyakanok, D. Roth, and W. Yih, "Natural language inference via dependency tree mapping: An application to question answering," *AI & Math*, January 2004.
[64] D. R. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, "Probabilistic question answering on the web," in *Proc. of the Int. WWW Conf.*, pp. 408–419, 2002.
[65] D. R. Radev, J. M. Prager, and V. Samn, "Ranking suspected answers to natural language questions using predictive annotation," in *Proceedings 6th*

Applied Natural Language Processing Conference (ANLP2000), pp. 150–157, Seattle, WA, 2000.

[66] D. R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan, and J. M. Prager, "Mining the web for answers to natural language questions," in *Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management*, Altlanta GA, 2001.

[67] D. Ravichandran and E. Hovy, "Learning surface text patterns for a question answering system," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL02)*, pp. 41–47, Philadelphia, July 2002.

[68] E. Rosch *et al.*, "Basic objects in natural categories," *Cognitive Psychology*, vol. 8, pp. 382–439, 1976.

[69] A. Singhal, J. Choi, D. Hindle, J. Hirschberg, F. Pereira, and W. Whittaker, "AT&T at TREC-7 SDR track," in *Proceedings of the Broadcast News Transcription and Understanding Workshop (BNTUW'099)*, 1999.

[70] M. Soubbotin, "Patterns of potential answer expressions as clues to the right answers," in *Proceedings of the 10th Text Retrieval Conference (TREC2001)*, pp. 293–302, NIST, Gaithersburg, MD, 2002.

[71] M. Soubbotin and S. Soubbotin, "Use of patterns for detection of answer strings: A systematic approach," in *Proceedings of the 11th Text Retrieval Conference (TREC2002)*, pp. 325–331, NIST, Gaithersburg, MD, 2003.

[72] T. Strzalkowski and S. Harabagiu, eds., *Advances in Open-Domain Question-Answering*, Springer, 2006.

[73] S. Tellex, B. Katz, J. Lin, G. Marton, and A. Fernandes, "Quantitative evaluation of passage retrieval algorithms for question answering," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pp. 41–47, Toronto, Canada, July 2003.

[74] E. Voorhees, "Query expansion using lexical-semantic relations," in *Proceedings of the Seventeenth International ACM-SIGIR Conference on Research and Development in Information Retrieval*, (W. Croft and C. van Rijsbergen, eds.), pp. 61–69, 1994.

[75] E. M. Voorhees and H. T. Dang, "Overview of the TREC 2005 question answering track," in *Proceedings of the 14th Text Retrieval Conference*, NIST, Gaithersburg, MD, 2006.

[76] E. M. Voorhees and D. K. Harman, eds., *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005.

[77] E. M. Voorhees and D. Tice, "Building a question answering test collection," in *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200–207, Athens, August 2000.

[78] N. Wacholder, Y. Ravin, and M. Choi, "Disambiguation of proper names in text," in *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC, April 1997.

[79] D. H. D. Warren and F. C. N. Pereira, "An efficient easily adaptable system for interpreting natural language queries," *Computational Linguistics*, vol. 8, no. 3–4, pp. 110–122, 1982.

[80] T. Winograd, "Procedures as a representation for data in a computer program for under-standing natural language," *Cognitive Psychology*, vol. 3, no. 1, 1972.
[81] W. A. Wood, "Progress in natural language understanding — An application to Lunar Geology," *AFIPS Conference Proceedings*, vol. 42, pp. 441–450, 1973.
[82] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4–11, Zurich, Switzerland, August 18–22 1996.
[83] J. Xu, A. Licuanan, and R. Weischedel, "TREC 2003QA at BBN: Answering definitional questions," in *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, Gaithersburg, MD, 2004.
[84] H. Yang, T. Chua, S. Wang, and C. Koh, "Structured use of external knowledge for event-based open domain question answering," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, 2003.