

Bibliography

I have not included any references to TREC proceedings papers in this bibliography. All of the TREC proceedings are available in the Publications section of the TREC web site. In addition to papers from individual participants, the proceedings contain track overview papers and an overview of the whole conference.

References

- [1] Pranav Anand, Eric Breck, Brianne Brown, Marc Light, Gideon Mann, Ellen Riloff, Mats Rooth, and Michael Thelen. Fun with reading comprehension. Final report of the Johns Hopkins 2000 Summer Workshop on Reading Comprehension. <http://www.clsp.jhu.edu/ws2000/groups/reading>.
- [2] The AnswerBus web site. <http://www.answerbus.com>.
Web-based, multilingual QA system.
- [3] D.E. Appelt, J.R. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. SRI International FASTUS system MUC-6 test results and analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 237–248. Morgan Kaufmann, 1995.
SRI’s named entity recognizer.
- [4] David Banks, Paul Over, and Nien-Fan Zhang. Blind men and elephants: Six approaches to TREC data. *Information Retrieval*, 1:7–34, 1999.
Statistical exploration of a set of TREC results. This is a good reference for the fact that the topic effect is bigger than the system effect in retrieval results.
- [5] BBN Systems and Technologies. BBN: Description of the PLUM system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 55–69. Morgan Kaufmann, 1995.
BBN’s named entity recognizer.
- [6] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 192–199, 2000.
Paper on finding answers from within a set of canned responses such as FAQ files or helpdesk applications.
- [7] Matthew W. Bilotti, Boris Katz, and Jimmy Lin. What works better for question answering: Stemming or morphological query expansion. In *IR4QA: Information Retrieval for Question Answering, A SIGIR 2004 Workshop*, pages 1–7, 2004.
Examination of how stemming affects retrieval results for QA. Demonstrates that the TREC QA collections are not reusable in that they are not complete.
- [8] Eric Breck, John Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light, and Inderjeet Mani. How to evaluate your question answering system every day . . . and still get real work done. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, volume 3, pages 1495–1500, 2000.

MITRE's word recall approach to automatic evaluation of question answering systems.

- [9] Chris Buckley. trec_eval IR evaluation package. Available from <ftp://ftp.cs.cornell.edu/pub/smart>.

Evaluation software used to score ranked retrieval runs in TREC.

- [10] Chris Buckley. Implementation of the SMART information retrieval system. Technical Report 85-686, Computer Science Department, Cornell University, Ithaca, New York, May 1985.

Description of the SMART retrieval system that is only slightly older than the version of SMART that is publicly available.

- [11] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In N. Belkin, P. Ingwersen, and M.K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

Experiments showing that some evaluation measures are inherently less stable than others. Also shows how stability is increased by using more topics.

- [12] Robin D. Burke, Kristian J. Hammond, Vladimir A. Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. Questions answering from frequently-asked question files: Experiences with the FAQ Finder system. Technical Report TR-97-05, The University of Chicago, Computer Science Department, June 1997.

Classic system for finding answers from FAQ lists intended for humans.

- [13] Nancy Chinchor, Lynette Hirschman, and David D. Lewis. Evaluating message understanding systems: An analysis of the third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–449, 1993.

Comprehensive evaluation of the MUC-3 tasks.

- [14] Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 358–365, 2001.

Paper that used the redundancy inherent in large data sets (the web) to validate the answers to factoid questions.

- [15] C. W. Cleverdon. The Cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192, 1967.

Experiments that were basis of Cranfield evaluation methodology.

- [16] Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr, David Gunning, and Murray Burke. The DARPA high-performance knowledge bases project. *AI Magazine*, pages 25–49, Winter 1998.

Summary of the HPKB project at its midpoint.

- [17] W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August 1998. ACM Press, New York.

- [18] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andre Ng. Web question answering: Is more always better? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298, 2002.

Microsoft paper that discusses using massive amounts of data (that are therefore likely to contain simple expressions of answers to factoid questions) rather than complicated patterns or deep processing.

[19] Dino Esposito. Talk to your data. Technical Report 322070, Microsoft Knowledge Base Article, August 1999. <http://msdn.microsoft.com/library>.

White paper on Microsoft's English Query database front end

[20] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

The book on WordNet.

[21] Junichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question answering challenge (QAC-1): Question answering evaluation at NTCIR workshop 3. In *Working Notes of the Third NTCIR Workshop Meeting, Part IV: Question Answering Challenge (QAC1)*, pages 1–10. National Institute of Informatics, Japan, 2002.

Overview of the first QA task within NTCIR.

[22] B.F. Green, A.K. Wolf, C. Chomsky, and K. Laughery. BASEBALL: An automatic question answerer. In *Proceedings of the Western Joint Computer Conference*, volume 19, pages 219–224, 1961. (Reprinted in *Readings in Natural Language Processing*, B.J. Grosz, K. Sparck-Jones and B.L. Webber, editors, Morgan Kaufmann, 1986).

Very early database front end.

[23] Sanda Harabagiu, John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weischedel. Issues, tasks, and program structures to roadmap research in question & answering (Q&A), October 2000. <http://www-nlpir.nist.gov/projects/duc/roadmapping.html>.

QA Roadmap document.

[24] Sanda Harabagiu, Dan Moldovan, Marius Paşca, Rada Mihalcea, Mihai Surdeanu, Răzvan Bunescu, Roxana Gîrju, Vasile Rus, and Paul Morărescu. The role of lexico-semantic feedback in open-domain textual question-answering. In *Proceedings of the Association for Computational Linguistics*, pages 274–281, July 2001.

Paper describing feedback loops in LCC QA system.

[25] L. Hirschman and R. Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4), 2001.

The introductory article to a special issue on question answering. Nice recap of the state of the field as of mid-2001.

[26] David Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pages 329–338, 1993.

Readable explanation of statistical tests that can be used to test the significance of differences in retrieval runs. A call for more such testing. Gives the assumptions of each test and suggests some ways to see if the tests are valid.

[27] Jakarta Lucene. <http://jakarta.apache.org/lucene/docs/index.html>, August 2004.

The home page for the Lucene system.

[28] Boris Katz. From sentence processing to information access on the world wide web. Paper presented at the AAAI Spring Symposium on Natural Language Processing for the World Wide Web, 1997. Electronic version at <http://www.ai.mit.edu/people/boris/webaccess>.

(Early) description of the MIT START system.

- [29] Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In Robert Korfage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 181–190, 1993.

Early system whose task was one of the inspirations for the first TREC QA track task.

- [30] Wendy Lehnert. A conceptual theory of question answering. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 158–164, 1977. (Reprinted in *Readings in Natural Language Processing*, B.J. Grosz, K. Sparck-Jones and B.L. Webber, editors, Morgan Kaufmann, 1986).

Reading comprehension work that focused on pragmatics (i.e., difference between an answer and a response).

- [31] Lemur: The lemur toolkit for language modeling and information retrieval. <http://www-2.cs.cmu.edu/~lemur/>, August 2004.

The home page for the Lemur system.

- [32] Bernardo Magnini, Mateo Negri, Roberto Prevete, and Hristo Tanev. Is it the right answer? Exploiting web redundancy for answer validation. In *Proceedings of ACL 2002*, pages 425–432, 2002.

Automatic tests for verifying candidate answer based on counting web pages that contain the candidate in the vicinity of question words.

- [33] George Miller. Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

Classic publication for WordNet.

- [34] Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu. Performance issues and error analysis in an open domain question answering system. In *Proceedings of ACL-2002*, pages 33–40, 2002.

Study of relative contribution of different feedback back loops within LCC QA system.

- [35] Dan I. Moldovan and Vasile Rus. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings ACL 2001*, pages 294–401, 2001.

Description of the logical forms and proof system used in the LCC QA system

- [36] Christof Monz. *From Document Retrieval to Question Answering*. Institute for Logic, Language and Computation; University of Amsterdam, 2003.

Doctoral thesis that explores the issue of how document retrieval effectiveness affects QA system effectiveness.

- [37] The Message Understanding Conference web site. http://www.itl.nist.gov/iaui/894.02/related_projects/muc.

- [38] John O’Connor. Answer-passage retrieval by text searching. *Journal of the American Society for Information Science*, pages 227–239, July 1980.

Very early passage retrieval paper.

- [39] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-answering by predictive annotation. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 184–191, July 2000.

The “predictive annotation” paper: IBM’s system that indexes documents with named entity tags so questions with those tags match when using traditional IR matching techniques.

- [40] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of ACL-2002*, pages 41–47, 2002.

Describes bootstrapping method for learning templates/patterns for particular question types.

- [41] I. Roberts and R. Gaizauskas. Evaluating passage retrieval approaches for question answering. In *Proceedings of the 26th European Conference on Information Retrieval*, pages 72–84, 2004.

Evaluation of different passage retrieval algorithms for question answering. Introduced the coverage and redundancy measures.

- [42] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

Classic reference for different interpretations of relevance.

- [43] ELF Software. Elf software home page. <http://www.elfsoftware.com/home.htm>.

Creators of the Access ELF database front end.

- [44] K. Sparck Jones. Reflections on TREC. *Information Processing and Management*, 31(3):291–314, 1995.

Assessment of TREC after TREC-2.

- [45] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an “ideal” information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

Paper describing how pooling could be used to build a large test collection.

- [46] Karen Sparck Jones. Further reflections on TREC. *Information Processing and Management*, 36(1):37–85, 2000.

Assessment of TREC updated after TREC-6.

- [47] Karen Sparck Jones and Peter Willett. Evaluation. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, chapter 4, pages 167–174. Morgan Kaufmann, 1997.

Introduction to the Evaluation section of the Readings book.

- [48] Karen Sparck Jones and Peter Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1997.

Reprints of classic papers in IR with commentary from the editors.

- [49] Alan Stuart. Kendall’s tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.

Definition of the Kendall tau correlation measure.

- [50] Jean Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28(4):467–490, 1992.

Exposition of the myriad different considerations in designing retrieval experiments.

- [51] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of SIGIR 2003 the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–47, 2003.

Quantitative evaluation of different passage retrieval algorithms for question answering

- [52] The Text REtrieval Conference web site. <http://trec.nist.gov>.

- [53] Ellen M. Voorhees. Special issue: The sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1), January 2000.
- Special issue of the journal *Information Processing and Management* devoted to TREC-6.
- [54] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.
- Study showing retrieval results are stable despite differences in relevance judgments.
- [55] Ellen M. Voorhees. The TREC question answering track. *Natural Language Engineering*, 4(7):361–378, 2001.
- Summary of the first two TREC QA tracks.
- [56] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002.
- Empirical determination of the error rate for comparing two retrieval systems as a function of the number of topics used and the size of the difference in scores.
- [57] Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, July 2000.
- NIST’s pattern approach to automatic evaluation of question answering systems.
- [58] B. Webber. Question answering. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 2, pages 814–822. Wiley, 1987.
- Early summary of work on question answering (from the natural language understanding perspective).
- [59] Terry Winograd. Five lectures on artificial intelligence. In A. Zampolli, editor, *Linguistic Structures Processing*, volume 5 of *Fundamental Studies in Computer Science*, pages 399–520. North Holland, 1977.
- Descriptions of the STUDENT and SHRDLU systems.
- [60] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes, 2nd edition*. Morgan Kaufmann, Inc. San Francisco, CA, 1999.
- Retrieval book. The MG retrieval system is based on this work.
- [61] W. A. Woods. Lunar rocks in natural English: Explorations in natural language question answering. In A. Zampolli, editor, *Linguistic Structures Processing*, volume 5 of *Fundamental Studies in Computer Science*, pages 521–569. North Holland, 1977.
- Description of the LUNAR system, which allowed geologists to ask questions about moon rocks returned from Apollo space missions. User tested at a geologist’s convention.
- [62] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In Croft et al. [17], pages 307–314.
- Investigation of the effect of pooling on evaluation. Shows that pool quality does matter, but TREC pools are more than sufficient to get reliable comparisons.
- [63] V. Zue, S. Seneff, J. Glass, J. Polofroni, C. Pao, T.J. Hazen, and L. Heatherington. JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):100–112, 2000.
- MIT weather information dialog system.