

Matthew W. Bilotti

Candidate for the Ph.D. in Language and Information Technologies at the Language Technologies Institute, Carnegie Mellon University, expected graduation December 2009.

Contact Information

Email: mbilotti@cs.cmu.edu
Phone: +1 412 260 8042
Web: <http://www.cs.cmu.edu/~mbilotti>

Research Interests

Information retrieval over semi-structured text annotated with linguistic and semantic content; machine learning approaches to ranking annotated text with respect to complex information needs; question answering.

Education

- 2004-present* Ph.D. in Language and Information Technologies, (expected December 2009).
Language Technologies Institute, School of Computer Science,
Carnegie Mellon University
Dissertation Title: Linguistic and Semantic Retrieval Strategies for Question Answering
Thesis Committee:
Eric Nyberg (Chair), Professor (ehn@cs.cmu.edu)
Jamie Callan, Professor (callan@cs.cmu.edu)
Jaime Carbonell, Professor (jgc@cs.cmu.edu)
Language Technologies Institute, Carnegie Mellon University
Eric Brown, Researcher (ewb@us.ibm.com)
IBM T.J. Watson Research Center
- 2003-2004* M.Eng. in Electrical Engineering and Computer Science, conferred June 2004.
Computer Science and Artificial Intelligence Laboratory
Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology
Thesis Title: Query Expansion Techniques for Question Answering
Thesis Supervisor: Boris Katz, Principal Research Scientist (boris@csail.mit.edu)
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
- 1999-2003* S.B. in Computer Science and Engineering, conferred June 2003.
Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology

Awards and Honors

- 2001 Elected Member, Eta Kappa Nu (www.hkn.org), the Electrical and Computer Engineering Honor Society, Beta Theta Chapter, Massachusetts Institute of Technology.
- 2002 Elected Member, Tau Beta Pi (www.tbp.org), the Engineering Honor Society, Mass Beta Chapter, Massachusetts Institute of Technology.
- 2002 Anthony Sun Fellowship for International Education
MIT International Science and Technology Initiatives
- 2003 Elected Associate Member, Sigma Xi (www.sigmaxi.org), the Scientific Research Society, Chapter 063, Massachusetts Institute of Technology.

Publications

- Matthew W. Bilotti, Jonathan Elsas, Jaime Carbonell and Eric Nyberg. Passage Ranking for Question Answering using Linguistic and Semantic Features. In Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009). 2009. (Submitted.)
- Matthew W. Bilotti and Eric Nyberg. Improving Text Retrieval Precision and Answer Accuracy in Question Answering Systems. In Proceedings of the Second Information Retrieval for Question Answering (IR4QA) Workshop at COLING 2008. 2008.
- Matthew W. Bilotti, Le Zhao, Jamie Callan and Eric Nyberg. Focused Retrieval over Richly-Annotated Collections. In Proceedings of the SIGIR 2008 Workshop on Focused Retrieval. 2008.
- Matthew W. Bilotti, Paul Ogilvie, Jamie Callan and Eric Nyberg. Structured Retrieval for Question Answering. In Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2007.
- Teruko Mitamura, Frank Lin, Hideki Shima, Mengqiu Wang, Jeongwoo Ko, Justin Betteridge, Matthew Bilotti, Andrew Schlaikjer and Eric Nyberg. JAVELIN III: Cross-Lingual Question Answering from Japanese and Chinese Documents. In Proceedings of NTCIR-6 Workshop. 2007.
- Matthew W. Bilotti and Eric Nyberg. Evaluation for Scenario Question Answering. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006). 2006. (Poster.).
- Eric Nyberg, Teruko Mitamura, Robert Frederking, Vasco Pedro, Matthew W. Bilotti, Andrew Schlaikjer and Kerry Hannan. Extending the JAVELIN QA System with Domain Semantics. In Proceedings of the Question Answering in Restricted Domains Workshop at AAAI 2005.
- Eric Nyberg, Robert Frederking, Teruko Mitamura, Matthew Bilotti, Kerry Hannan, Laurie Hiyakumoto, Jeongwoo Ko, Frank Lin, Lucian Lita, Vasco Pedro and Andrew Schlaikjer. JAVELIN I and II Systems at TREC 2005. In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005). 2005.
- Boris Katz, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Wesley Hildebrandt, Roni Katzir, Jimmy Lin, Daniel Loreto, Gregory Marton, Federico Mora and Ozlem Uzuner. Answering multiple questions on a topic from heterogeneous resources. In Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004). 2004.
- Matthew W. Bilotti, Boris Katz and Jimmy Lin. What Works Better for Question Answering: Stemming or Morphological Query Expansion?. In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004. 2004.
- Matthew W. Bilotti. Query Expansion Techniques for Question Answering. Master's thesis, Massachusetts Institute of Technology, 2004.
- Boris Katz, Jimmy Lin, Daniel Loreto, Wesley Hildebrandt, Matthew Bilotti, Sue Felshin, Aaron Fernandes, Gregory Marton and Federico Mora. Integrating Web-based and Corpus-based Techniques for Question Answering. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003). 2003.

Research Statement

If Question Answering (QA) systems are ever to reach the level of speed and accuracy required to be competitive with the web search engines that are ubiquitous in the lives of today's internet users, the quality of the underlying text retrieval process must be improved.

Most Information Retrieval (IR) systems, including those that are embedded in QA systems, are optimized to provide a quality ad hoc retrieval experience for a human user, but fail to address the unique needs of QA systems. QA systems often have a much more complete specification of what they are looking for than human users do. This specification consists of linguistic and semantic constraints that the system knows must hold for a piece of text to contain the answer to the question.

My dissertation research focuses on improving the quality of retrieved text within the context of a Question Answering (QA) system by applying learning-to-rank techniques to a feature space derived from the linguistic and semantic constraints of interest to the system.

The approach involves re-ranking the retrieval output from a bag-of-words and Named Entity baseline, which consists of keywords drawn from the question and a Named Entity type placeholder representing the expected answer type. This is considered to be a strong passage retrieval baseline for QA, because when the retrieval unit is small, it approximates density-based methods.

The trained linear model is used to re-rank the retrieved passages based the degree of partial satisfaction of the linguistic and semantic constraints derived from the question. Recent experiments with TREC QA data show that this method realizes significant improvements in Mean Average Precision with respect to a bag-of-words and Named Entity baseline.

Research Experience

*January 2008
to present*

Graduate Research Assistant
AKITA Project, Language Technologies Institute
Supervisor: Eric Nyberg, Professor, CMU

Applied rule-based inference techniques from the expert system literature to the task of assessing the health of a business. The project involved extracting business events from newswire and public filings, such as earnings reports, mergers and acquisitions and legal issues. Rules developed by a subject matter expert were used to aggregate evidence from the extracted content, and to predict whether or not a company was likely to be entering the downward spiral that precedes bankruptcy.

*June 2007
to June 2008*

Graduate Research Assistant
BlueJay Project, Language Technologies Institute
and the IBM T.J. Watson Research Center
Supervisor: Eric Nyberg, Professor, CMU

Participated in architecture design for Question Answering (QA) systems, focusing in particular on support for incorporating linguistic and semantic information into the retrieval process. The effort was intended to build a generalized QA system structure that supported the wide variety of architectural variants in use. System modules would become components in IBM's Unstructured Information Management Architecture (UIMA), and the framework would handle the information flow. The end goal was to enable sharing of components and comparative evaluation between QA research groups.

*August 2006
to December 2006*

Graduate Research Assistant
RADAR Project, Language Technologies Institute
Supervisor: Eric Nyberg, Professor, CMU

The RADAR project was a large, multi-disciplinary effort to build a desktop personal assistant able to assist a user with email management. As a part of a team building technologies for email message understanding, I built frameworks for batch linguistic and semantic annotation of text and persistence of annotated text. These technologies also found use as a part of the JAVELIN project, in which the important texts were not email messages, but a document collection over which a QA system would answer questions.

*August 2004
to June 2008*

Graduate Research Assistant
JAVELIN Project, Language Technologies Institute
Supervisor: Eric Nyberg, Professor, CMU

Performed research and development of Question Answering (QA) systems, with specific focus on retrieval strategies for QA and on text annotation and methods of exploiting annotations to improve retrieval performance. Early experiments focused on using query expansion and gradual relaxation strategies that dropped keywords if not enough documents were retrieved. Later experiments examined techniques for factoid QA over Chinese and Japanese corpora, and complex question answering over English newswire and in technical domains.

Recent experiments involve use of structured retrieval techniques and thoroughly-annotated corpora to improve retrieval performance for QA. Text annotation is done via the Unstructured Information Management Architecture (UIMA), with plug-ins that provide named entity recognition and verb predicate-argument structures with PropBank semantic role labels on the arguments. The Indri search engine is used to index the corpus; practically any UIMA type system can be indexed. These annotations are necessary for the system to distinguish an answer-bearing document from one that merely matches certain keywords. Structured queries encoding the linguistic and semantic constraints were shown to outperform a bag-of-words baseline over the AQUAINT corpus for 109 TREC factoid questions with exhaustive human relevance judgments. For QA tasks in other corpus languages and domains, and/or on other types of questions, a different type system may be required to get the best performance out of structured retrieval.

*January 2001 to
June 2004*

Undergraduate, and later Graduate Research Assistant
Infolab Group, MIT Computer Science and Artificial Intelligence Laboratory
Supervisor: Boris Katz, Principal Research Scientist, MIT

Conducted research and development of Question Answering (QA) technology, including knowledge-based approaches to complex QA and standard approaches to factoid QA, including answer projection. Researched query expansion techniques for retrieval for QA based on morphological and derivational alternations of query terms, in an effort to maximize retrieval of relevant documents. This query expansion technique used with the Lucene search engine provided a significantly higher-quality ranking of documents than the PRISE rankings provided by the Text Retrieval Conference (TREC) organizers in TREC 2003, though end-to-end system results did not improve significantly.

Industrial Experience

*December 2008
to March 2009* Retrieval Consultant
skribel, Inc., San Mateo, CA

Designed and implemented a variety of search strategies for skribel's application, which suggests web pages likely to be of interest to a user. Conducted performance evaluation using pooled relevance judgments and user studies. Authored proposals for submission to federal funding agencies. Java and C++ on a Linux platform, using technologies Tomcat, Spring, Lucene and the Lemur toolkit for Language Modeling and Information Retrieval (<http://www.lemurproject.org>).

Summer 2002 Software Engineer, Business Development Centre for Europe, Middle East, and Asia
IBM Italia S.p.A., via Lecco, 61, 20059 Vimercate (MI)

Designed, developed and deployed innovative software solutions for IBM, IBM Business Partners and the IBM customer base to facilitate commerce, customer support, licensing and contract management, software deployment and other tasks related to the IBM Business Model. ANSI C, Perl, Java and Web based technologies on IBM AIX and Microsoft Windows platforms.

Summer 1999, Software Engineer, Software Services Group for Rhode Island and Massachusetts
Summer 2000 GTECH Corporation, 55 Technology Way, West Greenwich, RI 02817

Maintained and managed deployments of proprietary database software on live server farms. ANSI C and DCL on a DEC OpenVMS platform.

Teaching Experience

Spring 2007, 11-792 Software Engineering for Information Systems II, CMU.
Spring 2008 Instructor: Eric Nyberg, Professor, Language Technologies Institute,
Carnegie Mellon University

In this course, students form small self-managing project teams that tackle interesting problems in language technologies research. I worked closely with each project group, advising them on both a technical level and a research level. There were three projects in the Spring of 2007: annotation, indexing and retrieval for business applications, semantic role labeling for nominalized predicates in a biomedical domain, and active learning for training annotators. In the Spring of 2008, a smaller class formed a single project group working on opinion mining over blog data. All three projects involve the use of the Unstructured Information Management Architecture (UIMA) to annotate text and the Indri search engine.

Fall 2006 11-791 Software Engineering for Information Systems I, CMU.
Instructor: Eric Nyberg, Professor, Language Technologies Institute,
Carnegie Mellon University

Developed assignments and exam questions, held weekly office hours and graded student problem sets, exams and projects. Presented several lectures, including one on design patterns. Familiarizing the students with UIMA was a specific educational objective of this course. To that end, I worked one-on-one with students having trouble with assignments designed specifically to force them to learn UIMA. I developed a course project based on an exercise in annotating text with UIMA, and met weekly with student project groups in the second half of the semester to advise and assess progress.

Spring 2004 6.871 Knowledge-Based Application Systems, MIT.
Instructors: Howard E. Shrobe, Principal Research Scientist, and
Kimberle Koile, Lecturer, Massachusetts Institute of Technology

Held weekly office hours, graded student problem sets, project reports, and reading summaries, which were due at the beginning of each lecture. Supported student project groups building expert systems with the Joshua truth maintenance package, using Lisp with the CLIM GUI toolkit.

Fall 2003

6.121J/HST.575J Bioelectronics Project Laboratory, MIT.

Instructor: Stephen Burns, Senior Research Scientist,
Massachusetts Institute of Technology

Taught 4, 2-hour lab sessions, held weekly office hours, graded student problem sets and project reports. Worked one-on-one with students assisting with circuit design and debugging and microcontroller programming. Held a tutorial at the beginning of the semester reviewing basic electronics and circuit analysis. Performed extensive curriculum development for, and supervised pilot project groups using, new Java-based microcontrollers.