

Curriculum Vitae

Grace Hui Yang

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA, USA, 15213

Telephone: (+1)412-2157651
Email: huiyang@cs.cmu.edu
Web: <http://www.cs.cmu.edu/~huiyang>

Research Interests

Information Retrieval, Text Mining, Machine Learning, Natural Language Processing, Human-Computer Interaction.

Education

- ❖ **Carnegie Mellon University**, Pittsburgh, PA, USA
Ph.D., Computer Science
Language Technologies Institute, School of Computer Science.
Advisor: Prof. Jamie Callan.
Thesis Topic: Personal Ontology Learning.
Thesis Committee: Jamie Callan, Jaime Carbonell, Christos Faloutsos, Ed Hovy.
Expected May 2011.
- ❖ **Carnegie Mellon University**, Pittsburgh, PA, USA
Master's in Language Technologies
Language Technologies Institute, School of Computer Science.
Advisor: Prof. Jamie Callan.
Project: Near-Duplicate Detection.
Aug 2004- May 2006.
- ❖ **National University of Singapore**, Singapore
Master's in Computer Science
Department of Computer Science, School of Computing.
Advisor: Prof. Tat-Seng Chua.
Thesis Topic: Event-based Question Answering.
Thesis Committee: Tat-Seng Chua, Hwee-Tou Ng, Min-Yan Kan.
Aug 2001 – May 2004.
- ❖ **National University of Singapore**, Singapore
B.Sc. in Computer Science (**First Class Honors**)
Department of Computer Science, School of Computing.
Mentor: Prof. Kim-Teng Lua.
Final Year Project: Online Video Recording and Streaming.
Aug 1997 – May 2001.

Professional Experience

- ❖ Research Assistant, Carnegie Mellon University, USA
Aug 2006 – present
Worked on personal ontology learning for the eRulemaking project. The work investigates human-guided machine learning methods of creation of light-weight, personal ontologies that allow users to quickly understand the range of the issues raised, and enable “drilling down” into documents that discuss a particular topic.
- ❖ Research Intern, Microsoft Research/Bing, Redmond, WA, USA
Jun–Sep 2009
Mentor: Anton Mityagin, Manager: Rich Caruana.

Worked on search engine training and evaluation. It explores whether, when, and for which data points one should obtain multiple, expert training labels, as well as what to do with the labels once they have been obtained. Collecting multiple overlapping labels only for a subset of training samples that has already been labeled relevant is far more effective.

- ❖ Research Assistant, Carnegie Mellon University, USA
Aug 2004 – Aug 2006
Worked on Near-Duplicate Detection for the eRulemaking project. The work develops a near-duplicate detection algorithm that flexibly incorporates instance-level constraints into semi-supervised clustering.
- ❖ Research Assistant, National University of Singapore, Singapore
July-Aug 2004
Worked on news video retrieval and question answering. The TREC QA Track attempts to deal with open-domain factoid, list, and definitional questions. The event-based question answering approach exploits general ontologies and external resources, such as WordNet glosses and synonyms, and search result snippets, to gather additional world knowledge about a question-answer event, in which the answer lies.
- ❖ Full-time Teaching Assistant, National University of Singapore, Singapore
2001- 2004
Conducted 10-12 hour tutorials per week, managed course materials for the classes of software engineering, multimedia technologies, and artificial intelligence.
- ❖ System Engineer, National University of Singapore, Singapore
May-July 2001
Worked on mobile services via short messaging services (SMS).
- ❖ System Analyst, GlobalID Asia, Singapore
May-July 2001
Worked on RFID tracking systems for logistics and animal tagging.
- ❖ Engineering Intern, Singapore Telecommunication, Singapore
May 1999 – Jan 2000
Developed tools for designing telecommunication services.

Teaching Experience

- ❖ Teaching Assistant, Carnegie Mellon University, Pittsburgh, PA, USA
Search Engine and Web Mining (Undergraduate) F10
Information Retrieval (Graduate) So8
- ❖ Full-time Teaching Assistant, National University of Singapore, Singapore
Software Engineering (Undergraduate) Fo1, Fo2, So3, Fo3, So4
Artificial Intelligence (Undergraduate) Fo3
Multimedia Technologies (Undergraduate) So2
- ❖ Teaching Assistant, National University of Singapore, Singapore
Programming in Java (Undergraduate) Foo, So1

Honors& Awards

- ❖ Facebook Fellowship finalist (2010; 1 out of 22 finalists nationwide)
- ❖ Google SIGIR 2010 conference grant award (2010; 1 out of 2 award winners)
- ❖ 2nd position Question Answering system in TREC 2002, TREC 2003 and TREC 2004.
- ❖ 1st position Video Retrieval system in TRECVID 2003 and TRECVID 2004.
- ❖ First Class Honors in Computer Science, National University of Singapore (2001; top 1%)
- ❖ Dean's List, National University of Singapore (2000, top 1%)
- ❖ Singapore Ministry of Education Scholarship (1996-2001; 1996's top 200 students in China)
- ❖ Excellent Youth Award in Province, China (1995)
- ❖ 1st Prize in Province, National Mathematics Olympics Competition, China (1993)
- ❖ 2nd Prize in Province, National Chemistry Olympics Competition, China (1993)
- ❖ 1st Prize in Province, Essay Competition, China (1992)

Publications

❖ Journal Articles

Hui Yang and Jamie Callan. OntoCop: Constructing Ontologies for Public Comments. *Journal of IEEE Intelligent Systems*, 24(5), page 70-75. IEEE Computer Society. 2009.

❖ Refereed Full Papers

Hui Yang, Anton Mityagin, Krysta Svore, and Sergey Markov. Collecting High Quality Overlapping Labels at Low Cost. In *Proceedings of the 33th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR2010), Geneva, Switzerland. July 19-23, 2010.

[Acceptance rate: 16.7%]

Hui Yang and Jamie Callan. A Metric-based Framework for Automatic Taxonomy Induction. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics* (ACL2009), Singapore. Aug 2-7, 2009.

[Acceptance rate: 21%]

Hui Yang and Jamie Callan. Ontology generation for large email collections. In *Proceedings of the 8th National Conference on Digital Government Research* (Dg.O 2008). Montreal, Canada. May, 2008.

Hui Yang and Jamie Callan. Near-Duplicate Detection by Instance-level Constrained Clustering, In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR2006), Seattle, WA. Aug 6-11 2006.

[Acceptance rate: 18.5%]

Hui Yang, Jamie Callan, Stuart Shulman, Next Steps in Near-Duplicate Detection for eRulemaking. In *Proceedings of the 6th National Conference on Digital Government Research* (Dg.O2006), San Diego, CA. May 21-24 2006.

Hui Yang, Jamie Callan. Near-Duplicate Detection for eRulemaking, In *Proceedings of the 5th National Conference on Digital Government Research* (Dg.O2005), Atlanta, GA. 15-18 May 2005.

Hui Yang, Tat-Seng Chua. Web-based List Question Answering. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004), Geneva, Switzerland. 23-27 Aug 2004.

Hui Yang, Tat-Seng Chua, Shuguang Wang, Chun-Keat 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* (SIGIR 2003), Toronto, Canada. 28 July-1 Aug 2003.

[Acceptance rate: 17%]

Hui Yang, Lekha Chaison, Yunlong Zhao, Shi-Yong Neo, Tat-Seng Chua. VideoQA: Question Answering on News Video. In *Proceedings of the 11th Annual ACM International Conference on Multimedia* (ACM MM 2003), Berkeley, CA. Nov 2-8, 2003. [Acceptance rate: 17%]

Hui Yang, Tat-Seng Chua. QUALIFIER: Question Answering by Lexical Fabric and External Resource. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2003), Budapest, Hungary. April 12-17, 2003.

[Acceptance rate: 26%]

❖ Refereed Workshop Papers, Posters, and Demonstrations

Hui Yang and Jamie Callan. Feature Selection for Automatic Taxonomy Induction. (Poster) In *Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR2009), Boston, MA. July 19-23, 2009. [Poster acceptance rate: 33.6%]

Hui Yang and Jamie Callan. Metric-Based Ontology Learning. Workshop on *Ontologies and Information Systems for the Semantic Web of 17th Conference on Information and Knowledge Management* (CIKM2008). Napa Valley, CA. Oct, 2008.

Hui Yang and Jamie Callan. Learning the Distance Metric in a Personal Ontology. Workshop on *Ontologies and Information Systems for the Semantic Web of 17th Conference on Information and Knowledge Management* (CIKM2008). Napa Valley, CA. Oct, 2008.

Hui Yang and Jamie Callan. Human-Guided Ontology Learning. In *Second Workshop on Human-Computer Interaction and Information Retrieval* (HCIR2008). Microsoft Research, Redmond. Oct, 2008.

Hui Yang, Jamie Callan, Stuart Shulman. DURIAN: A demo for near-duplicate detection (demo description). In *Proceedings of the 6th National Conference on Digital Government Research* (Dg.O2006), San Diego, CA. May 21-24, 2006.

Hui Yang, Tat-Seng Chua. Effective Web Page Classification for Finding List Answers. (poster) In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2004), Sheffield, UK. Aug 25-29, 2004.

Hui Yang, Tat-Seng Chua. FAD on the Web: Find All Distinct Answers. (poster) In *Proceedings of the 13th International World Wide Web Conference* (WWW 2004), New York. 17-22 May 2004.

Hui Yang, Tat-Seng Chua, Shuguang Wang. Modeling Web Knowledge for Answering Event-based Questions. (poster) In *Proceedings of the 12th International World Wide Web Conference* (WWW 2003), Budapest, Hungary. May 2003.

❖ Non-Refereed Publications

Hui Yang, Luo Si, Jamie Callan. Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track. In *Proceedings of the 15th Text REtrieval Conference 2006* (TREC2006), Gaithersburg, MD. Nov 14-17 2006.

Alex Hauptmann, Ming-Yu Chen, Mike Christel, C. Huang, W.-H. Lin, Toubin Ng, N. Papernick, A. Velivelli, Jun Yang, Rong Yan, **Hui Yang**, and H.D. Wactlar. Confounded Expectations: Informedia at TRECVID 2004. In *the 13th Text REtrieval Conference Video Workshop* (TRECVID2004), Gaithersburg, MD. Nov 15-16 2004.

Tat-Seng Chua, Yunlong Zhao, Lekha Chaisorn, Chun-Keat Koh, **Hui Yang**, Huaxin Xu, Qi Tian. TREC 2003 Video Retrieval and Story Segmentation Task at NUS PRIS. In *the 12th Text REtrieval Conference Video Workshop* (TRECVID2003), Gaithersburg, MD. Nov 2003.

Hui Yang, Hang Cui, Min-Yen Kan, Mstislav Maslennikov, Long Qiu, Tat-Seng Chua. QUALIFIER in TREC-12 QA Main Task. In *Proceedings of the 12th Text REtrieval Conference* (TREC'2003), Gaithersburg, MD. Nov 2003.

Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. In *Proceedings of the 11th Text REtrieval Conference* (TREC2002), Gaithersburg, MD. Nov 19-22 2002.

Presentations

❖ Invited Talks

Question Answering and VideoQA. Georgia Institute of Technology. Nov 10, 2003.

Metric-based Ontology Learning. Chinese Academy of Science. Aug 15, 2008.

Introduction to Ontology Learning. Xi'an Jiaotong University. Sep 5, 2008.

Personal Ontology Learning. Georgetown University. March, 2011.

Personal Ontology Learning. Massachusetts Institute of Technology. March, 2011.

❖ TREC Plenary Talks

The Integration of Lexical Knowledge and External Resources for Question Answering. TREC 2002 QA Track. Nov 19 2002.

QUALIFIER in TREC-12 QA Main Task. TREC 2003 QA Track. Nov 2003.

❖ Conference and Workshop Talks

Collecting High Quality Overlapping Labels at Low Cost. SIGIR2010.

Metric-based Ontology Learning. ACL 2009.

Human-Guided Ontology Learning. HCIR2008.

Learning the Distance Metric in a Personal Ontology. CIKM2008.

Metric-Based Ontology Learning. CIKM2008.

Ontology generation for large email collections. Dg.O2008.

Near-Duplicate Detection by Instance-level Constrained Clustering. SIGIR2006.

Next Steps in Near-Duplicate Detection for eRulemaking. Dg.O2006.

Near-Duplicate Detection for eRulemaking. Dg.O2005.

VideoQA: Question Answering on News Video. ACM Multimedia 2003.

Structured Use of External Knowledge for Event-based Open Domain Question Answering. SIGIR2003.

QUALIFIER: Question Answering by Lexical Fabric and External Resources. EACL2003.

Professional Services

❖ **Program Committee**

- The 34th International ACM SIGIR Conference on Research and Development for Information Retrieval. SIGIR Poster Program Committee 2011.
- The 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010.
- The 10th Annual International Conference on Digital Government Research. Dg.O 2009.

❖ **Organizing Committee**

- Chair for Student Research Track, the 10th Annual International Conference on Digital Government Research. Dg.O 2009.
- Co-organizer for the Information Retrieval Discussion Series, Carnegie Mellon University. 2006-present.

❖ **Reviewer**

- The 27th International ACM SIGIR Conference on Research and Development for Information Retrieval. SIGIR 2004.
- The 28th International ACM SIGIR Conference on Research and Development for Information Retrieval. SIGIR 2005.
- The 30th International ACM SIGIR Conference on Research and Development for Information Retrieval. SIGIR 2007.
- The 31st International ACM SIGIR Conference on Research and Development for Information Retrieval, SIGIR 2008.

Industrial Collaborations

❖ **Spin-off Company**

Telltale Information, Pittsburgh, PA, USA.

❖ **IT Company**

Microsoft Bing, Redmond, WA, USA.

Language Skills

- ❖ English (fluent in both speaking and writing)
- ❖ Chinese (native)

Relevant Graduate Coursework (Selected)

- ❖ Statistical Machine Learning, by Larry Wasserman & John Lafferty. Spring 2009.
- ❖ Optimization, by Geoff Gordon & Carlos Guestrin. Spring 2008.
- ❖ Intermediate Statistics, by Matthew Harrison. Fall 2007.
- ❖ Advanced IR Seminar, by Jamie Callan & Yiming Yang. Fall 2007.
- ❖ Machine Learning, by Carlos Guestrin. Spring 2007.
- ❖ Advanced Machine Learning Seminar, by Yiming Yang. Fall 2006.
- ❖ Grammar Formalism, by Lori Levin. Spring 2006.
- ❖ Computer-Mediated Communication, by Susan Fussell, Spring 2006.
- ❖ Software Engineering, by Eric Nyberg. Fall 2005.
- ❖ Information Retrieval, by Jamie Callan & Yiming Yang. Spring 2005.
- ❖ Personalized Information Retrieval, by Joemon Jose. Spring 2005.
- ❖ Language & Statistics, by Roni Rosenfeld. Spring 2005.
- ❖ Grammar & Lexicon, by Lori Levin. Fall 2004.
- ❖ Algorithms for NLP, by Alon Lavie & Robert Frederking. Fall 2004.
- ❖ Hypermedia Information Processing, by Tat-Seng Chua. Spring 2001.