DAVID ARTHUR SMITH

home:
87 Butterfield Terrace
Amherst, MA 01002, USA
Phone: +1 413 549 9650
Mobile: +1 410 900 6238

office:
Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003-9264, USA
Phone: +1 413 545 1381

online:
Email: dasmith@cs.umass.edu
http://www.cs.umass.edu/~dasmith

Education

Johns Hopkins University 2010 **Ph.D. in Computer Science**
Advisor: Jason Eisner
- *National Science Foundation fellowship (2003–6)*
- *Wolman fellowship (2002–3)*

Harvard University 1994 **A.B. *summa cum laude* in Classics (Greek)**
- *Harvard National Scholar*

Professional Experience

Northeastern University September 2012 –
Assistant Professor, College of Computer and Information Science

University of Massachusetts Amherst September 2008 –
Research Assistant Professor, Department of Computer Science, Center for Intelligent Information Retrieval

Johns Hopkins University September 2002 – September 2008
Research Assistant, Department of Computer Science, Center for Language and Speech Processing
Machine learning for natural language processing: semi-supervised learning and efficient inference techniques; syntactic parsing; morphological disambiguation; machine translation and word alignment
Summer Research Workshop, 2003: Member of Syntax for Statistical Machine Translation team

Google, Inc. May 2005 – September 2005
Internship in Machine Translation group
Research on improved training and decoding for machine translation

Tufts University July 1994 – August 2002
Perseus Digital Library Project
Information retrieval and extraction, named-entity disambiguation, digital libraries, document layout analysis, document alignment, morphological analysis

Teaching Experience

Search Engines Spring 2012
Department of Computer Science, University of Massachusetts Amherst (446)
Instructor for undergraduate course on information retrieval; enrollment: 40

Freshman Computer Science Seminar Fall 2011
Department of Computer Science, University of Massachusetts Amherst (191a)
Co-instructor for Residential Academic Program Seminar; enrollment: 24

Introduction to Natural Language Processing Fall 2009
Department of Computer Science, University of Massachusetts Amherst (585)

Designer and Instructor
Advanced undergraduate/graduate class with students from computer science and linguistics; enrollment: 17

Mining Text and Images in Digital Libraries Using Grid Computing Spring 2009
Department of Computer Science, University of Massachusetts Amherst (791MT)
Designer and Instructor, with James Allan and R. Manmatha
Graduate seminar with readings and final project; enrollment: 10

Empirical Research Methods in Computer Science Fall 2005
Department of Computer Science, Johns Hopkins University (600.408)
Designer and Primary Instructor (with Noah Smith)
One-credit course for advanced undergraduates and graduate students on computer-intensive statistics and experimental design; enrollment: 18

An Overview of Statistical Machine Translation August 2006
Conference of the Association for Machine Translation in the Americas, Cambridge, MA
Designer and Primary Instructor (with Charles Schafer)
Tutorial on data, models, and algorithms in statistical MT for broad audience; enrollment: 12

Invited course lectures:
Tufts University (CS 0150-TC, Classics 0191-TC), Information retrieval in digital libraries, February 2002

Grants and Contracts

DARPA BOLT: *Effective Interactive Retrieval Combining Multiple Annotations and Representations* (co-PI, $2.8M) 2011–2016

DARPA Machine Reading: *A Universal Machine Reading System* (co-PI, $2.5M) 2009–2014

NSF Data-Intensive Computing: *Mining a Million Books: Linguistic and Structure Analysis, Fast Expanded Search, and Improved OCR* (co-PI, $2.3M) 2009–2013

Army/MURI: *SUBTLE: Situation Understanding Bot through Language and Environment* (co-PI, $634k) 2007–2012

NIH Clinical and Translational Science program (CIIR membership subcontract, $40k) 2010–2015

NSF CluE: *Learning Word Relationships Using TupleFlow* (senior personnel, $450k) 2009–2011

Yahoo!, Inc.: *Data-Intensive Processing for Better Search, Analysis, and OCR* (PI, in-kind access to Yahoo!'s Hadoop cluster) 2009–2011

NEH Start-up Grants: *OCRonym: Entity Extraction and Retrieval for Scanned Books* (co-PI, $50k) 2009–2010

Dissertation

[1] David A. Smith. *Efficient Inference for Trees and Alignments: Modeling Monolingual and Bilingual Syntax with Hard and Soft Constraints and Latent Variables*. PhD thesis, Johns Hopkins University, 2010.

Refereed Conference Proceedings

[2] Jason Naradowsky, Sebastian Riedel, and David A. Smith. Improving NLP through maginalization of hidden syntactic structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

[3] Sebastian Riedel, David A. Smith, and Andrew McCallum. Parse, price and cut—delayed column and row generation for graph based parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012.

[4] Yanchuan Sim, Noah A. Smith, and David A. Smith. Discovering factions in the computational linguistics community. In *ACL Workshop on Rediscovering 50 Years of Discoveries*, 2012.

[5] Michael Bendersky and David A. Smith. A dictionary of wisdom and wit: Learning to extract quotable phrases. In *NAACL Workshop on Computational Linguistics for Literature*, 2012.

[6] Jeffrey Dalton, James Allan, and David A. Smith. Passage retrieval for incorporating global dependencies in sequence labeling. In *Conference on Information and Knowledge Management (CIKM)*, pages 355–364, 2011.

[7] Jinyoung Kim, W. Bruce Croft, David A. Smith, and Anton Bakalov. Evaluating an associative browsing model for personal information. In *Conference on Information and Knowledge Management (CIKM)*, pages 647–652, 2011.

[8] Jae-Hyun Park, W. Bruce Croft, and David A. Smith. A quasi-synchronous dependence model for information retrieval. In *Conference on Information and Knowledge Management (CIKM)*, pages 17–26, 2011.

[9] David A. Smith, R. Manmatha, and James Allan. Mining relational structure from millions of books: Position paper. In *Proceedings of the CIKM BooksOnline Workshop*, pages 49–54, 2011.

[10] Kriste Krstovski and David A. Smith. A minimally supervised approach for detecting and ranking document translation pairs. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 207–216, 2011.

[11] Michael Bendersky, W. Bruce Croft, and David A. Smith. Joint annotation of search queries. In *Proceedings of the Association for Computational Linguistics*, pages 102–111, 2011.

[12] John S. Y. Lee, Jason Naradowsky, and David A. Smith. A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the Association for Computational Linguistics*, pages 885–894, 2011.

[13] Elif Aktolga, James Allan, and David A. Smith. Passage reranking for question answering using syntactic structures and answer types. In *European Conference on Information Retrieval (ECIR)*, pages 617–628, 2011.

[14] Michael Bendersky, W. Bruce Croft, and David A. Smith. Structural annotation of search queries using pseudo-relevance feedback. In *Conference on Information and Knowledge Management (CIKM)*, pages 1537–1540, 2010.

[15] Jinyoung Kim, Anton Bakalov, David A. Smith, and W. Bruce Croft. Building and evaluating a semantic representation for personal information. In *Conference on Information and Knowledge Management (CIKM)*, pages 1741–1744, 2010.

[16] Xiaobing Xue, W. Bruce Croft, and David A. Smith. Query reformulation using query distributions. In *Conference on Information and Knowledge Management (CIKM)*, pages 1497–1500, 2010.

[17] Sebastian Riedel, David A. Smith, and Andrew McCallum. Inference by minimizing size, divergence, or their sum. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 227–234, 2010.

[18] Sebastian Riedel and David A. Smith. Relaxed marginal inference and its application to dependency parsing. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 760–768, 2010.

[19] Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using thread structure. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 1907–1910, 2009.

[20] David A. Smith and Jason Eisner. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 822–831, 2009.

[21] David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–889, 2009.

[22] Michael Bendersky, W. Bruce Croft, and David A. Smith. Two-stage query segmentation for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference*, pages 810–811, 2009.

[23] David A. Smith and Jason Eisner. Dependency parsing by belief propagation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 145–156, 2008.

[24] David A. Smith and Jason Eisner. Bootstrapping feature-rich dependency parsers with entropic priors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 667–677, 2007.

[25] David A. Smith and Noah A. Smith. Probabilistic models of nonprojective dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 132–140, 2007.

[26] Keith Hall, Jiří Havelka, and David A. Smith. Log-linear models of non-projective trees, k-best MST parsing and tree-ranking. In *Proceedings of the CoNLL Shared Task*, pages 962–966, 2007.

[27] David A. Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics*, pages 787–794, 2006.

[28] David A. Smith and Jason Eisner. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30, 2006.

[29] Markus Dreyer, David A. Smith, and Noah A. Smith. Vine parsing and minimum risk reranking for speed and precision. In *Proceedings of the CoNLL Shared Task*, pages 201–205, 2006.

[30] Noah A. Smith, David A. Smith, and Roy W. Tromble. Context-based morphological disambiguation with random fields. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 475–482, 2005.

[31] David A. Smith and Noah A. Smith. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 49–56, 2004.

[32] F.J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A smorgasbord of features for statistical machine translation. In *Proceedings of the Conference on Human Language Technology and the North American Association for Computational Linguistics*, pages 161–168, 2004.

[33] David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References*, pages 45–49, 2003.

[34] Gregory Crane, Clifford E. Wulfman, Lisa M. Cerrato, Anne Mahoney, Thomas L. Milbank, David Mimno, Jeffrey A. Rydberg-Cox, David A. Smith, and Christopher York. Towards a cultural heritage digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2003*, pages 75–86, Houston, TX, June 2003.

[35] David A. Smith. Detecting and browsing events in unstructured text. In *Proceedings of the 25th Annual ACM SIGIR Conference*, pages 73–80, Tampere, Finland, August 2002.

[36] David A. Smith. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 191–196, Portland, OR, July 2002.

[37] David A. Smith, Anne Mahoney, and Gregory Crane. Integrating harvesting into digital library content. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 183–184, Portland, OR, July 2002.

[38] Gregory Crane, David A. Smith, and Clifford E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, June 2001. **Best paper award**.

[39] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of the European Conference on Digital Libraries (ECDL)*, pages 127–136, Darmstadt, Germany, September 2001.

[40] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. In *Proceedings of Extreme Markup Languages 2000*, pages 219–224, Montreal, August 2000.

Refereed Journal Articles

[41] David Bamman and David A. Smith. Extracting two thousand years of Latin from a million book library. *ACM Journal on Computing and Cultural Heritage*, 5(1), 2012.

[42] Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using conversational structures. *Information Retrieval*, 14(6):547–571, 2011.

[43] Andrew Kae, David A. Smith, and Erik Learned-Miller. Learning on the fly: A font-free approach towards multilingual OCR. *International Journal on Document Analysis and Recognition*, 14(3):289–301, 2011.

[44] Gregory R. Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Drudgery and deep thought: Designing a digital library for the humanities. *Communications of the Association for Computing Machinery*, 44(5):35–40, 2001.

[45] David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[46] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. *Markup Languages: Theory and Practice*, 2(3):205–214, 2000.

[47] David A. Smith. Textual variation and version control in the TEI. *Computers and the Humanities*, 33(1-2):103–112, 1999.

Other Publications

[48] Xiaoye Wu and David A. Smith. Right-branching tree transformation for eager dependency parsing. Technical Report CIIR-776, University of Massachusetts, 2010.

[49] Jason Naradowsky, Joe Pater, David Smith, and Robert Staubs. Learning hidden metrical structure with a log-linear model of grammar. In *Computational Modelling of Sound Pattern Acquisition*, pages 59–60, Edmonton, February 2010. Department of Linguistics, University of Alberta.

[50] Joe Pater, David A. Smith, Robert Staubs, Karen Jesney, and Ramgopal Mettu. Learning hidden structure with a log-linear model of grammar. In *Linguistic Society of America (LSA)*, Baltimore, January 2010.

[51] Gregory Druck and David A. Smith. Computing conditional feature covariance under non-projective tree conditional random fields. Technical Report UM-CS-2009-060, University of Massachusetts, 2009.

[52] David A. Smith. Debabelizing libraries: Machine translation by and for digital collections. *D-Lib Magazine*, 12(3), March 2006.

[53] Anne Mahoney, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman. Generalizing the Perseus XML document manager. In *Linguistic Exploration: Workshop on Web-based Language Documentation and Description*, Philadelphia, December 2000.

Invited Presentations

- Virginia Tech, Computer Science Department, March 2012
- CUNY, Graduate Center, March 2012
- Carnegie Mellon University, Language Technologies Institute, March 2012
- University of Chicago, Computer Science Department, February 2012
- Toyota Technical Institute, Chicago, February 2012
- Syracuse University, School of Information Studies, February 2012
- Yale University, Linguistics Department, February 2012
- Cornell University, AI Seminar, February 2012
- Northeastern University, College of Computer and Information Science, February 2012
- Harvard University, Institute for Quantitative Social Science, May 2011
- Princeton University, Computer Science Department, February 2011
- Carnegie Mellon University, Language Technologies Institute, February 2011
- MIT, Computer Science and Artificial Intelligence Laboratory, January 2011
- Humboldt University, Berlin, Institut für deutsche Sprache und Linguistik, January 2011
- UCLA, Institute for Pure and Applied Mathematics, August 2010
- University of Edinburgh, School of Informatics, March 2008
- University of Pittsburgh, Computer Science Department, February 2008
- University of Maryland, Computer Science Department, February 2008

Advising

Doctoral Committees

- Kedar Bellare. Advisor, Andrew McCallum. 2009 (proposal) –
- Gregory Druck. Advisor, Andrew McCallum. 2009 – 2011 (defended)
- David Mimno. Advisor, Andrew McCallum. 2009 – 2011 (defended)
- Lisa Friedland. Advisor, David Jensen. 2010 (proposal) –
- Jangwon Seo. Advisor, Bruce Croft. 2010 – 2011 (defended)
- Xiaobing Xue. Advisor, Bruce Croft. 2010 (proposal) –
- Michael Bendersky. Advisor, Bruce Croft. 2010 (proposal) –
- Jinyoung Kim. Advisor, Bruce Croft. 2011 (proposal) –

Current Advisees

- Jason Naradowsky. UMass Ph.D. student; co-advisor, Andrew McCallum. 2008–
- Kriste Krstovski. UMass Ph.D. student. 2009–
- Xiaoye Wu. UMass Ph.D. student. 2009–

Other Research Supervised

- Elif Aktolga (UMass Ph.D. student). Qualifying synthesis project, with James Allan. 2009.
- Andrew Kae (UMass Ph.D. student). Qualifying synthesis project, with Erik Learned-Miller. 2009–10.
- Jacqueline Feild (UMass Ph.D. student). Qualifying synthesis project, with Erik Learned-Miller. 2009–10.
- David Goff (Cornell undergraduate). Summer REU Site advisee. 2010.
- Jeff Dalton (UMass Ph.D. student). Qualifying synthesis project, with James Allan. 2010–11.

Service

Journal reviewing: *Computational Linguistics*, *Computers and the Humanities*, *Literary and Linguistic Computing*, *Proceedings of the National Academy of Sciences*

Conference reviewing: ACH/ALLC, ACL, COLING (Machine Learning area chair), DH10, ICML, HLT-NAACL, EACL, EMNLP, IJCNLP (Machine Learning area chair), SIGIR

Department committees: graduate program committee (UMass, 2010–11); ad-hoc committee for new institute for computational and experimental linguistics (UMass, 2009–10); curriculum (UMass, 2008–10); graduate student recruiting (JHU, 2003–7), system administration (JHU, 2003–8)

Software

Programmer for document management system for the **Perseus Digital Library** (http://www.perseus.tufts.edu) 1999–2002. One of the largest, and most popular, humanities digital libraries, Perseus presents sources for language, literature, art, and archaeology for several periods from the ancient Mediterranean through 19th century North America. Users viewing documents receive automatically generated information on morphology, lexicon, translations, technical terms, and named entities, as well as temporal and spatial visualizations.

Programmer for **Perseus: Sources and Studies on Ancient Greece**, 2.0 (Yale U. P., 1996), 3.0 (Yale U. P., 2000).

Personal Details

Date of Birth: 27 October 1972

Citizenship: USA

Languages: English (native); ancient Greek, Latin, French, German (reading); Arabic (basic)