DAVID ARTHUR SMITH

home:
4611 Arabia Avenue
Baltimore, MD 21214, USA
Phone: +1 410 254 3569
Mobile: +1 410 900 6238

office:
Johns Hopkins University
224 New Engineering Building
3400 North Charles Street
Baltimore, MD 21218, USA
Phone: +1 410 516 6238

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

EDUCATION

Johns Hopkins University	2008 (expected)	**Ph.D. in Computer Science** *Thesis:* Parser bootstrapping and cross-language projection *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*

RESEARCH EXPERIENCE

Johns Hopkins University September 2002 – present
Department of Computer Science, Center for Language and Speech Processing
Machine learning for natural language processing: semi-supervised techniques and numerical optimization; 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

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 of 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 of 12

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

REFEREED CONFERENCE PROCEEDINGS

Natural Language Processing and Machine Learning

[1] 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. **Nominated for best paper award**.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, pages 49–56, 2004.

[9] 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.

Information Extraction and Retrieval

[10] 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.

[11] 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.

[12] 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.

[13] 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.

Digital Libraries

[14] 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.

[15] 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.

[16] 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**.

[17] 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

[18] David A. Smith, Jason Eisner, and Noah A. Smith. Minimum risk annealing: Case studies in nonprojective dependency parsing and machine translation. Submitted to *Computational Linguistics*, 2008.

[19] 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.

[20] 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.

[21] 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.

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

Other Publications

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

[24] 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

- Humboldt University, Berlin, Linguistics Department, March 2008
- Imperial College, London, Internet Centre, March 2008
- University of Edinburgh, School of Informatics, March 2008
- University of Massachusetts, Amherst, Computer Science Department, March 2008
- University of Pittsburgh, Computer Science Department, February 2008
- University of Maryland, Computer Science Department, February 2008
- Tufts University, Computer Science Department, December 2007
- Carnegie Mellon University, Language Technologies Institute, February 2007

SERVICE

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

Conference reviewing: ACL, HLT-NAACL, EMNLP, IJCNLP, ACH/ALLC, Digital Humanities

Department committees: graduate student recruiting (Center for Language and Speech Processing, 2003–7), system administration (2003–8)

SOFTWARE

Programmer for document management system for the **Perseus Digital Library** (http://www.perseus.tufts.edu) 1999–2002. One of the largest heterogeneous 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. As of fall 2005, traffic on this site had reached 15,000,000 page views to 500,000 users a month.

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)