

SAJIB DASGUPTA

150 Palm Valley Road, Apt 2068, San Jose, CA 95123

Phone: 214-718-0541, Email: sdgnew@gmail.com

EDUCATION

Ph.D., Computer Science

University of Texas at Dallas (2005-2010)

Advisor: Vincent Ng

GPA: 4.0 (out of 4.0)

B.Sc., Computer Science & Engineering

Bangladesh University of Engineering and Technology (1999-2004)

GPA: 3.85 (out of 4.0)

RESEARCH INTERESTS

Natural Language Processing, Machine Learning, Data Mining. Special interests in natural language processing are unsupervised learning, sentiment classification, language-independent text processing.

PROFESSIONAL EXPERIENCE

Postdoctoral Researcher (2010 – Current)

IBM Almaden Research Center

San Jose, CA

Worked on SystemT project, a declarative information extraction system, which provides scalable knowledge mining from unstructured text data.

Research Assistant (2006 – 2010)

Human Language Technology Research Institute (HLTRI)

University of Texas at Dallas

Researched on multi-faceted text clustering which is aimed at generating diverse clusterings of a document collection according to user interests. Researched on automatic review classification and language-independent morphological word segmentation, with an application to language-independent part-of-speech induction.

Research Engineer (2007 – 2008)

IBM Almaden Research Center

San Jose, CA

Worked on the MatchMaker project, where the goal was to learn cross-corpus associations from unstructured datasets with an aim to bridging the gap between two disparate subject areas.

Lecturer and Research Programmer (2004 – 2005)

Department of Computer Science and Engineering

Center for Research on Bangla Language Processing (CRBLP)

BRAC University, Bangladesh

Taught Artificial Intelligence at the undergraduate level. Researched on knowledge-driven two level morphological parsing of Bengali and morphology-driven text processing, which was partially supported by IDRC, Canada.

SELECTED PUBLICATIONS

Text and Sentiment Classification

- Towards Subjectifying Text Clustering. *Sajib Dasgupta* and Vincent Ng. Accepted for presentation in the conference of ACM Special Interest Group on Information Retrieval (**SIGIR**), Geneva, 2010.
- Mining Clustering Dimensions. *Sajib Dasgupta* and Vincent Ng. Accepted for presentation in the conference of International Conference on Machine Learning (**ICML**), Haifa, 2010.
- Which Clustering do you Want? Inducing your Ideal Clustering using Minimal Feedback. *Sajib Dasgupta* and Vincent Ng. Accepted for publication in the Journal of Artificial Intelligence Research (**JAIR**), 2010.
- Single Data, Multiple Clusterings. Accepted for presentation in the **NIPS workshop** on Clustering, Vancouver, 2009.
- Topic-wise, Sentiment-wise, or Otherwise? Identifying the Hidden Dimension for Unsupervised Text Classification. In the conference of the Empirical Methods in Natural Language Processing (**EMNLP**), Singapore, 2009.
- Mine the Easy, Classify the Hard: Experiments with Automatic Sentiment Classification. *Sajib Dasgupta* and Vincent Ng. In the 47th Annual Meeting of the Association for Computational Linguistics (**ACL**), Singapore, 2009.
- Examining the Role of Linguistics Knowledge Sources in the Automatic Identification and Classification of Reviews. Vincent Ng, *Sajib Dasgupta* and S. M. Niaz Arifin. In the 44th Annual Meeting of the Association for Computational Linguistics (**ACL**), Sydney, 2006.

Unsupervised Learning of Linguistic Knowledge

- Unsupervised Part-of-Speech Acquisition for Resource-Scarce Languages. *Sajib Dasgupta* and Vincent Ng. In the conference of the Empirical Methods in Natural Language Processing (**EMNLP**), Prague, 2007.
- High-Performance, Language-Independent Morphological Segmentation. *Sajib Dasgupta* and Vincent Ng. In the annual conference of the North American Chapter of the Association for Computational Linguistics (**NAACL**), New York, 2007.
- Unsupervised Morphological Parsing of Bengali. *Sajib Dasgupta* and Vincent Ng. In the journal of Language Resources and Evaluation, 2007.
- Unsupervised Word Segmentation for Bangla. *Sajib Dasgupta* and Vincent Ng. In the International Conference on Natural Language Processing (**ICON**), 2007.

Machine Translation

- An Optimal Way of Machine Translation from English to Bengali. *Sajib Dasgupta*, Abu Wasif and Sharmin Azam. In the conference of International Conference on Computer and Information Technology (**ICCIT**), Bangladesh, 2004.

Invited Position Paper

- Discriminative Models for Semi-supervised Natural Language Learning. *Sajib Dasgupta*, Vincent Ng. Position paper in the NAACL-HLT 2009 workshop on Semisupervised Learning for Natural Language Processing, Boulder, 2009.

PATENT

Information Extraction from Multiple Expertise-Specific Subject Areas. Submitted by IBM. Docket no. ARC920080067US1. With co-inventors Dipayan Gangopadhyay and Norm Pass.

IMPORTANT PROJECTS

- Text clustering with interactive feedback (Matlab).
- Unsupervised word segmentation/stemming for English, Finnish, Turkish and Bengali (C++).
- Language independent part-of-speech acquisition for English and Bengali (C++).
- Sentiment classification in a supervised and semi-supervised setting (Matlab, C++).
- Unsupervised association rule mining from unstructured text with an application to projects/assets searching (Java).
- Automatic thesaurus construction from raw corpus (Java, Perl).
- Knowledge-based word stemming and interactive search and replace (C++, Java).
- Machine Translation from English to Bengali using CYK parsing (C++).

THESES

PhD: Multi-faceted Clustering: Spectral and Probabilistic Approaches.

Supervisor: Vincent Ng, University of Texas at Dallas, 2010 (Expected).

Masters: Toward Language Independent Morphological Segmentation and Part-of-speech Induction. Supervisor: Vincent Ng, University of Texas at Dallas, 2007.

ACADEMIC ACHIEVEMENTS

- *Louis Beecherl Jr. Graduate Fellowship*, University of Texas at Dallas, 2009-2010.
- *Graduate Student Scholarship*, University of Texas at Dallas, four academic years.
- SIGIR Student Travel Award for the conference of SIGIR, 2010, Geneva.
- Student Travel Scholarship for NIPS workshop on clustering, 2010, Geneva.
- *Dean's List* scholarship, Bangladesh University of Engineering and Technology, four academic years.
- *Talent Pool Scholarship* (the highest ranked scholarship given by the Education Board of Controller, Bangladesh), four academic years.
- 2nd in the Higher Secondary School Certificate examination, Chittagong, Bangladesh.

PROFESSIONAL SERVICE

Program Committee (Reviewer):

IJCAI 2011, EMNLP 2011, EMNLP 2010, COLING 2010, ACL-IJCNLP 2009, EMNLP 2008

TALKS

- Main Conference of SIGIR, 2010, Geneva.
- IBM T. J. Watson Research Center, 2010, New York.
- NIPS workshop on Clustering, 2009, Vancouver.
- Main Conference of ACL, 2009, Singapore.
- Main Conference of EMNLP, 2009, Singapore.
- Main Conference of NAACL, 2007, New York.