

Curriculum Vitae

Anoop Sarkar

School of Computing Science
8888 University Drive
Burnaby, BC V5A 1S6, Canada

Email: anoop@cs.sfu.ca

Web: <http://www.cs.sfu.ca/~anoop>

Tel.: +1 (778) 782 4933

Fax: +1 (778) 782 3045

Contents

1	Background	2
1.1	Education	2
1.2	Employment History	2
1.3	Awards	2
2	Research	3
2.1	Research Program	3
2.2	Most Significant Research Contributions	3
2.3	Publications	4
2.4	Research Experience	9
2.5	Invited Talks and Presentations	10
3	Service	11
3.1	Service to the Academic Community	11

1 Background

1.1 Education

- 2002 Ph.D. in Computer Science
Department of Computer and Information Sciences,
University of Pennsylvania, USA
Thesis: Combining Labeled and Unlabeled Data in Statistical Natural Language Parsing
Advisor: Prof. Aravind Joshi
My thesis was among the first to introduce semi-supervised learning techniques to the area of natural language parsing. It also introduced novel statistical parsing algorithms for Tree-adjoining grammars.
- 1991 B.Eng. in Computer Science
University of Poona, India
First Class with Distinction (ranked third overall)

1.2 Employment History

- 09/2009 - current Associate Professor
School of Computing Science, Simon Fraser University
- 09/2002 - 09/2009 Assistant Professor
School of Computing Science, Simon Fraser University
- 06/2002 - 09/2002 Postdoctoral Fellow
Institute for Research in Cognitive Science (IRCS), University of Pennsylvania
- 08/1995 - 06/2002 Research Assistant
Department of Computer and Information Sciences, University of Pennsylvania
- 08/1991 - 08/1993 Research Associate
Center for Development in Advanced Computing (C-DAC), Poona, India

1.3 Awards

- 2011 NSERC Collaborative Research and Development (CRD) Grant, \$350,000
Type: Research
Organization: NSERC
http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/CRD-RDC_eng.asp
- 2009 Google Faculty Research Award, \$70,000
Type: Research
Organization: Google Inc.
http://research.google.com/university/research_awards.html
- 2009 Best Paper Award
Type: Research
Conference: Canadian AI 2009
- 2008 IBM Faculty Award, \$20,000
Type: Research
Organization: International Business Machines (IBM) Inc.
<http://www-304.ibm.com/jct09002c/university/scholars/facultyawards/index.html>
- 2007 Excellence in Undergraduate Teaching Award

Type: Teaching

Organization: SFU Undergraduate Computing Science Student Society (CSSS)

1993-1998 Dean's Fellowship

Type: Fellowship

Organization: University of Pennsylvania

2 Research

2.1 Research Program

My research focuses on *machine learning* algorithms applied to the study of natural language. I am especially interested in unsupervised and *semi-supervised learning* for *natural language processing* (NLP). In contrast to the study of binary classifiers common in machine learning, my work in such algorithms is related to the study of structured outputs like sequences of tags, or parse trees. The significance of semi-supervised learning to NLP and to machine learning can be seen by its many benefits: (a) the expensive annotation of training data can be minimized, or if training data is unlabeled then small amounts of annotation can be used to improve performance; (b) NLP tools can be built for resource-poor languages where large amounts of training data are scarce; and (c) humans seem to learn language in a semi-supervised manner, so this research can provide insight into how humans acquire language.

I have also made several contributions to the study of *stochastic tree-adjoining grammars*, a generalization of context-free grammars which allows for a computationally constrained and linguistically sophisticated analysis of natural language. In particular, I focus on formal properties of stochastic grammars, the implementation of statistical parsers that can be trained to parse natural language, and grammar induction from text corpora. More recently, I have explored the connections between tree-adjoining grammar and tree automata, tree transducers and synchronous grammars, and their application to natural language semantics and machine translation.

My research also focuses on the applications of *statistical parsing* to various natural language processing tasks such as *machine translation*, multi-document query-based summarization, and information extraction from bio-medical and newspaper texts. In each case, I aim to show that basic research into statistical parsing can benefit the performance on these varied tasks.

2.2 Most Significant Research Contributions

Conference publications are extremely important in my research area. My publication record in important conferences, in particular the meetings of the Association for Computational Linguistics (ACL), is substantive, and the overall number is among the strongest compared to peers in this research area at other Canadian universities.

The paper numbers in this section refer to the publications in Section 2.3.

Semi-supervised learning for parsing natural language

My Ph.D. thesis work was one of the first to explore semi-supervised learning in parsing natural language and led to subsequent joint work which was funded as part of the NSF funded Johns Hopkins research workshop series. My tutorial, Paper #60, (co-authored with my Ph.D. student, Gholamreza Haffari) on semi-supervised learning is a popular tutorial on the topic. Along with my students, we have published several notable publications in this research area.

Parsing natural language using Tree-adjoining grammars

Tree-adjoining grammar (TAG) is a framework that has attractive properties as a formalism to represent natural language syntax. I have made several contributions to the development and implementation of parsing algorithms for this framework. I am the developer of a widely used open-source software for parsing using TAG, which is about 20,000 lines of code in C, C++, Perl and Java. It has been used in several institutions around the world, including

the University of Pennsylvania (where the project originated), Rutgers University, DFKI in Saarbrücken, the Tsujii Research Lab at the University of Tokyo, and the University of Paris 7, among others. The software is available for download at: <http://www.cis.upenn.edu/~xtag/swrelease.html>

Recent work with my Ph.D. student, Yudong Liu, uses TAG for the semantic role labeling task, which is useful for the analysis of the underlying predicate argument structure of language. We have shown that using latent TAG-based features using Latent Support Vector Machines can do significantly better, by several percentage points, compared to the previous best method on this task.

Statistical machine translation

Statistical machine translation (MT) often relies on additional monolingual data from the target language, for example additional English sentences for Chinese–English translation. Paper #19, a collaboration with the National Research Council (NRC, Gatineau) was the first to show that by using semi-supervised learning, additional monolingual data from the source language can also improve translation quality. This makes it possible to translate languages or domains in which data is scarce. This result was used by NRC to improve their standing in the National Institute of Standards (NIST) 2006 MT competition.

Most of the field of MT focuses on translation into English. In recent work, I have instead focused on machine translation into morphologically complex languages such as Finnish, Czech, Turkish, German, etc. – which is far less studied. Our initial work has already attracted funding from Google Inc.

2.3 Publications

Legend

* Single author paper *or* multiple authors where I was primary author

† Multiple authors – equal contribution with other primary authors

‡ Multiple authors – I was not primary author

Names in **bold** face are my students.

Journal Papers

1. * A. Sarkar. Book Review of *Parsing Schemata for Practical Text Analysis* by Carlos Gómez Rodríguez. *Computational Linguistics* (to appear). 2011. doi: 10.1162/COLLR_00081
2. † N. Ueffing, **G. Haffari** and A. Sarkar. Transductive learning for statistical machine translation. In *Machine Translation*, DOI 10.1007/s10590-008-9036-3, Springer. 2008.
3. † K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi and B. Webber. D-LTAG System: Discourse parsing with a Lexicalized Tree Adjoining Grammar. *Journal of Logic, Language and Information: Volume 12, Issue 3*, pp. 261-279, Springer. 2003.
4. † S. Wintner and A. Sarkar. A note on typing feature structures. *Computational Linguistics*. 28(3):389-397. MIT Press. 2002.

Book Chapters

5. * A. Sarkar. Syntax and Parsing. In *Multilingual Natural Language Applications: From Theory to Practice*. Daniel M. Bikel and Imed Zitouni, eds. Prentice Hall. to appear in 2011.
6. † N. Ueffing, **G. Haffari** and A. Sarkar. Semi-supervised learning for machine translation. Book chapter in *Learning Machine Translation*, Eds. C. Goutte, N. Cancedda, M. Dymetman and G. Foster. MIT Press. 2008.
7. * A. Sarkar. Combining SuperTagging with Lexicalized Tree-Adjoining Grammar parsing. Book chapter in *Complexity of Lexical Descriptions and its Relevance to Natural Language Processing: A Supertagging Approach*, Eds. S. Bangalore and A. Joshi. 2007.

8. * A. Sarkar and A. Joshi. Tree-Adjoining Grammars and its application to statistical parsing. In *Data-oriented parsing*. Eds. R. Bod, R. Scha and K. Sima'an, CSLI Publications, Stanford, 2003.
9. † C. Doran, B. A. Hockey, A. Sarkar, B. Srinivas and F. Xia. Evolution of the XTAG System. In *Tree Adjoining Grammars: Formal, Computational and Linguistic Aspects*. pp. 371-404. Eds. A. Abeille and O. Rambow, CSLI Publications, Stanford, 2000.
10. * A. Sarkar. The conflict between future tense and modality: the case of will in English. In *Penn Working Papers in Linguistics* volume 5, number 2, pp. 91-117, 1998.

Refereed Conference Papers

11. † **Ann Clifton** and Anoop Sarkar. Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, OR, USA. June 19-24, 2011.
12. † Gholamreza Haffari, **Marzieh Razavi** and Anoop Sarkar. An Ensemble Model that Combines Syntactic and Semantic Clustering for Discriminative Dependency Parsing. Short Paper (poster). In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, OR, USA. June 19-24, 2011.
13. † **D. Song** and A. Sarkar. Training Global Linear Models for Chinese Word Segmentation. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009*. Kelowna, BC. May 25-27, 2009. **Best paper award.**
14. † **G. Haffari**, M. Roy and A. Sarkar. Active Learning for Statistical Phrase-based Machine Translation. In *Proceedings of the annual meeting of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*. Boulder, Colorado. May 31-June 5, 2009.
15. † **G. Haffari** and A. Sarkar. Active Learning for Multilingual Statistical Machine Translation. In *Proceedings of the 47th annual meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*. Singapore, August 2-7, 2009.
16. ‡ **G. Haffari** and A. Sarkar. Homotopy-based Semi-Supervised Hidden Markov Models for Sequence Labeling. In *Proc. of the 22nd International Conference on Computational Linguistics: COLING 2008*. Manchester, 18-22 August, 2008.
17. ‡ **G. Melli**, M. Ester and A. Sarkar. Recognition of Multi-sentence n-ary Subcellular Localization Mentions in Biomedical Abstracts. In *Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM) 2007*. Singapore, Dec 6-7, 2007.
18. † **G. Haffari** and A. Sarkar. Analysis of semi-supervised learning with the Yarowsky algorithm. In *Proc. of the 23rd Conf. on Uncertainty in Artificial Intelligence, UAI 2007*. Vancouver, BC. July 19-22, 2007.
19. † N. Ueffing, **G. Haffari** and A. Sarkar. Transductive learning for statistical machine translation. In *Proc. of the Annual Conf. of the Assoc. for Computational Linguistics, ACL 2007*, Prague, Czech Republic. June 25-27, 2007.
20. † **Y. Liu** and A. Sarkar. Experimental evaluation of LTAG-based features for Semantic Role Labeling. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the Conf. on Computational Natural Language Learning, EMNLP-CoNLL 2007*. Prague, Czech Republic. June 28-30, 2007.
21. † **Z. Shi**, **G. Melli**, Y. Wang, **Y. Liu**, B. Gu, **M. K. Kashani**, A. Sarkar and F. Popowich. Question answering summarization of multiple biomedical documents. In *Proc. of the 20th Canadian Conf. on Artificial Intelligence, Canadian AI 2007*, Montreal, QC. May 28-30, 2007.
22. † **Z. Shi**, A. Sarkar and F. Popowich. Simultaneous identification of biomedical named-entity and functional relations using statistical parsing techniques. In *Proc. of the Annual Conf. of the North American Chapter of the Assoc. for Computational Linguistics, NAACL-HLT 2007*, short paper, Rochester, USA. April 22-27, 2007.

23. † **J. Birke** and A. Sarkar. A clustering approach for the nearly unsupervised recognition of non-literal language. In *Proc. of the 11th Conf. of the European Chapter of the Assoc. for Computational Linguistics*, EACL-2006. Trento, Italy. April 3-7, 2006.
24. † **Z. Shi** and A. Sarkar. Intimate learning: A novel approach for combining labeled and unlabeled data. In Poster Track, *Nineteenth International Joint Conf. on Artificial Intelligence: IJCAI-05*. Edinburgh, UK. August 2-5, 2005.
25. † **H. Shen** and A. Sarkar. Voting between multiple data representations for text chunking. In *Proc. of the Eighteenth Meeting of the Canadian Society for Computational Intelligence*, Canadian AI 2005. Victoria, BC, Canada. May 9-11, 2005. In *Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 3501*, eds. Balász Kégl and Guy Lapalme. Springer.
26. † L. Shen, A. Sarkar and F. Och. Discriminative re-ranking for machine translation. In the *Human Language Tech. Conf. and the 5th Meeting of the North American Assoc. for Computational Linguistics: HLT-NAACL 2004*. Boston, USA. May 2-7, 2004.
27. † F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, D. Radev. A smorgasbord of features for statistical machine translation. In the *Human Language Tech. Conf. and the 5th Meeting of the North American Assoc. for Computational Linguistics: HLT-NAACL 2004*. Boston, USA. May 2-7, 2004.
28. † L. Shen, A. Sarkar and A. Joshi. Using LTAG-based features in parse re-ranking. In the *2003 Conf. on Empirical Methods in Natural Language Processing*. Sapporo, Japan. July 11-12, 2003.
29. † M. Steedman, R. Hwa, Stephen Clark, M. Osborne, A. Sarkar, J. Hockenmaier, P. Ruhlen, S. Baker, J. Crim. Example selection for bootstrapping statistical parsers. In the *Human Language Tech. Conf. and the 4th Meeting of the North American Assoc. for Computational Linguistics: HLT-NAACL 2003*. Edmonton, AB. May 27-June 1, 2003.
30. † M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, J. Crim. Bootstrapping statistical parsers from small data-sets. In *Proc. of the 11th Conf. of the European Assoc. for Computational Linguistics: EACL 2003*. Budapest, Hungary. April 12-17, 2003.
31. * A. Sarkar and W. Tripasai. Learning verb argument structure from minimally annotated corpora. In *Proc. of the 18th International Conf. on Computational Linguistics: COLING 2002*. Taipei, Taiwan. August 2002.
32. * A. Sarkar. Applying co-training methods to statistical parsing. In *Proc. of the 2nd Meeting of the North American Chapter of the Assoc. for Computational Linguistics: NAACL 2001*. pp. 175-182. Pittsburgh, USA, June 2001.
33. † A. Sarkar and D. Zeman. Automatic Extraction of Subcategorization Frames for Czech. In *Proc. of the 18th International Conf. on Computational Linguistics: COLING 2000*. pp. 691-698. Saarbrücken, Germany, August 2000.
34. † D. Zeman and A. Sarkar. Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In *2nd International Conf. on Language Resources and Evaluation: LREC 2000*. pp. 227-233. Athens, Greece. May 31 - June 2, 2000.
35. * A. Sarkar. Conditions on Consistency of Probabilistic Tree Adjoining Grammars. In *Proc. of the 36th Annual Meeting of the Assoc. for Computational Linguistics and 17th International Conf. on Computational Linguistics: COLING-ACL 1998*. pp. 1164-1170. Montreal, Quebec, 1998.
36. † M.-J. Nederhof, A. Sarkar and G. Satta. Prefix Probabilities from Probabilistic Tree Adjoining Grammars. In *Proc. of the 36th Annual Meeting of the Assoc. for Computational Linguistics and 17th International Conf. on Computational Linguistics: COLING-ACL 1998*. pp. 953-959. Montreal, Quebec, 1998.
37. * A. Sarkar. Separating Dependency from Constituency in a Tree Rewriting System. In *Proc. of the Fifth Meeting on Mathematics of Language*. pp. 153-160. Saarbrücken, Germany, August 1997.

38. † A. Sarkar and A. Joshi. Coordination in Tree Adjoining Grammars: Formalization and Implementation. In *Proc. of 16th International Conf. on Computational Linguistics: COLING 1996*. pp. 610-615. Copenhagen, Denmark, 1996.
39. * A. Sarkar. Incremental Parser Generation for Tree Adjoining Grammars. In *Proc. of the 34th Meeting of the Assoc. for Computational Linguistics: ACL 1996, Student Session*. pp. 375-376. University of California, Santa Cruz, June 1996.
40. * A. Sarkar. Extending Kimmo's Two-Level Model. In *Proc. of the 31st Meeting of the ACL, Student Session*. pp. 304-306. Columbus, Ohio. June 1993.

Refereed Workshop Papers

41. † **B. Sankaran, G. Haffari** and Anoop Sarkar. Bayesian Extraction of Minimal SCFG Rules for Hierarchical Phrase-based Translation. Sixth Workshop on Statistical Machine Translation: WMT-2011. Edinburgh, UK July 30-31, 2011.
42. † **B. Sankaran, A. Grewal** and Anoop Sarkar. Incremental Decoding for Phrase-based Statistical Machine Translation. (Full Paper). The Pacific Northwest Regional NLP Workshop (NW-NLP 2010). April 23, 2010.
43. † **A. Clifton** and Anoop Sarkar. Morphology Generation for Statistical Machine Translation Using Conditional Random Fields. (Full Paper). The Pacific Northwest Regional NLP Workshop (NW-NLP 2010). April 23, 2010.
44. † **Y. Liu, G. Haffari** and Anoop Sarkar. Latent SVMs for Semantic Role Labeling using LTAG Derivation Trees. (Full Paper). The Pacific Northwest Regional NLP Workshop (NW-NLP 2010). April 23, 2010. (**Best Presentation Award**)
45. **B. Sankaran, A. Grewal** and A. Sarkar. Incremental Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Uppsala, Sweden. July 15, 2010.
46. † **Y. Liu** and A. Sarkar. Exploration of the LTAG-Spinal Formalism and Treebank for Semantic Role Labeling. In *Grammar Engineering Across Frameworks (GEAF 2009)*. Workshop at the ACL/IJCNLP 2009 Conference. Singapore, August 6, 2009.
47. ‡ **M. Kashani**, F. Popowich, and A. Sarkar. Automatic transliteration of proper nouns from Arabic to English. In *Proc. of the Second Workshop on Computational Approaches to Arabic Script-based Languages, CAASL-2*. LSA 2007 Linguistic Institute, Stanford University. July 21-22, 2007.
48. † **Y. Liu, Z. Shi** and A. Sarkar. Exploiting rich syntactic information for relation extraction from bio-medical articles. In *Proc. of the Annual Conf. of the North American Chapter of the Assoc. for Computational Linguistics*, NAACL-HLT 2007, poster track, Rochester, USA. April 22-27, 2007.
49. † **J. Birke** and A. Sarkar. Active learning for the identification of non-literal language. In *Proc. of the Workshop on Computational Approaches to Figurative Language*, NAACL-HLT 2007 workshop, Rochester, USA. April 26, 2007.
50. † **Y. Liu** and A. Sarkar. Using LTAG-based features for Semantic Role Labeling. *Proc. of the Eighth Workshop on Tree Adjoining Grammars and Related Formalisms: TAG+8*, Poster Track, COLING-ACL 2006 workshop, Sydney, Australia. July 15-16, 2006.
51. † R. Hwa, M. Osborne, A. Sarkar and M. Steedman. Corrected co-training for statistical parsers. In *Proc. of the ICML Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining at the 20th International Conf. on Machine Learning (ICML-2003)*. Washington DC, USA. August 21-24, 2003.
52. † A. Sarkar and C. Han. Statistical morphological tagging and parsing of Korean with an LTAG grammar. In *Proc. of the Sixth Workshop on Tree Adjoining Grammars*. Venice, Italy. May 20-23, 2002.

53. † K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi and B. Webber. D-LTAG System - Discourse parsing with a Lexicalized Tree Adjoining Grammar. In *Information Structure, Discourse Structure and Discourse Semantics* Workshop, ESSLI 2001.
54. * A. Sarkar, F. Xia and A. Joshi. Some Experiments on Indicators of Parsing Complexity for Lexicalized Grammars. In *Efficiency in Large-Scale Parsing Systems* Workshop held at the 18th International Conf. on Computational Linguistics: COLING 2000. Luxembourg, 5 August 2000.
55. † R. Prasad and A. Sarkar. Comparing Test-suite based evaluation and Corpus-based evaluation of a wide-coverage grammar for English. In *Using Evaluation within HLT Programs: Results and Trends* LREC'2000 Satellite Workshop. Athens, Greece. May 30, 2000.
56. * A. Sarkar. Practical Experiments in Parsing using Tree Adjoining Grammars. In *Proc. of the Fifth Workshop on Tree Adjoining Grammars*, TAG+ 5. Paris, France. May 25-27, 2000.
57. † M.-J. Nederhof, A. Sarkar and G. Satta. Prefix Probabilities from Linear Indexed Grammars. In *Proc. of the Fourth Workshop on Tree Adjoining Grammars*, TAG+ 4, Philadelphia, August 1998.
58. † B. Srinivas, A. Sarkar, C. Doran and B. A. Hockey. Grammar and Parser Evaluation in the XTAG Project. Workshop on *Evaluation of Parsing Systems*, Granada, Spain, 26 May 1998.
59. ‡ C. Doran, B. Hockey, P. Hopely, J. Rosenzweig, A. Sarkar, B. Srinivas, F. Xia, A. Nasr and O. Rambow. Maintaining the Forest and Burning out the Underbrush in XTAG. Workshop on *Computational Environments for Grammar Development and Language Engineering* (ENVGRAM), Madrid, Spain, July 1997.

Tutorials

60. * Tutorial on inductive semi-supervised learning methods: with applicability to natural language processing. A. Sarkar and **G. Haffari**. Tutorial at the *Human Language Tech. Conf. - North American chapter of the Assoc. for Computational Linguistics* annual meeting (HLT-NAACL) 2006. New York City, USA. June 4, 2006. (*peer reviewed*)
61. * Parsing with Tree-adjoining Grammars. A. Sarkar. Tutorial at the *ACL/HCSNet Advanced Program in Natural Language Processing*, Melbourne, Australia. July 13-14, 2006. (*invited*)
62. * Chunking and Statistical Parsing. A. Sarkar. Tutorial at a two-week course on corpus-based NLP at Anna University, Chennai, India. December, 2001. (*invited*)

Non-refereed Workshop Publications

63. † **D. Song** and A. Sarkar. Training a perceptron with global and local features for Chinese word segmentation. Bake-off short paper. Proc. of the Sixth SIGHAN Workshop on Chinese Language Processing, IJCNLP 2008 workshop, Hyderabad, India. January 11-12, 2008.
64. † **D. Song** and A. Sarkar. Voting between dictionary-based and sub-word tagging models for Chinese word segmentation. Bake-off short paper. Proc. of the *Fifth SIGHAN Workshop on Chinese Language Processing*, COLING-ACL 2006 workshop, Sydney, Australia. July 22-23, 2006.
65. † **G. Melli**, **Z. Shi**, Y. Wang, **Y. Liu**, A. Sarkar and F. Popowich. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2006 summarization task. In *Proc. of the Document Understanding Conf. 2006* (DUC-2006), New York City, USA, June 8-9, 2006.
66. † **Z. Shi**, B. Gu, F. Popowich and A. Sarkar. Synonym-based query expansion and Boosting-based re-ranking: A two-phase approach for genomic information retrieval. In Proc. of the Fourteenth Text REtrieval Conf. (TREC 2005). NIST, Gaithersburg, USA, November 15-18, 2005.
67. † **G. Melli**, Y. Wang, **Y. Liu**, **M. Kashani**, **Z. Shi**, B. Gu, A. Sarkar and F. Popowich. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 summarization task. *Proc. of the Document Understanding Conf. 2005* (DUC-2005), Vancouver, B.C., October 2005.
68. ‡ B. Baldwin, J.C. Reynar, M. Collins, J. Eisner, A. Ratnaparkhi, J. Rosenzweig, A. Sarkar and B. Srinivas. Description of the University of Pennsylvania entry in the MUC-6 competition. *Proc. Sixth Message Understanding Conf.* (MUC-6), Maryland, October 1995.

Technical Reports

69. † Experimental evaluation of LTAG-based features for semantic role labeling. **Y. Liu** and A. Sarkar. Technical Report TR 2007-03, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, April 2007. (longer version of our EMNLP-CoNLL 2007 paper)
70. ‡ Analysis of semi-supervised learning with the Yarowsky algorithm. **G. Haffari** and A. Sarkar. Technical Report TR 2007-07, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, March 2007. (longer version of our UAI 2007 paper)
71. † F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, D. Radev. CLSP Final Report of the Johns Hopkins Summer Workshop 2003: Syntax for Statistical Machine Translation. 2003.
72. † M. Steedman, S. Clark, R. Hwa, M. Osborne, A. Sarkar, J. Hockenmaier, P. Ruhlen, S. Baker, J. Crim. CLSP Final Report of the Johns Hopkins Summer Workshop 2002: Semi-Supervised Training for Statistical Parsing. 2002.
73. † A Lexicalized Tree Adjoining Grammar for English. XTAG Research Group. Technical Report IRCS-01-03, IRCS, University of Pennsylvania, 2001.

2.4 Research Experience

74. Dec–Jun 2006, Document Understanding Conference (DUC) 2006
Team Leader, Question Answering Combined with Multi-Document Summarization.
The task was similar to the DUC 2005 effort (see below).
75. Jan–Jun 2005, Document Understanding Conference (DUC) 2005
Team Leader, Question Answering Combined with Multi-Document Summarization.
The task involved providing a 250 word summary answer to a provided question. The answer was produced from a small set of documents (20-50 per question). NIST ran the evaluation effort for all competing systems.
76. Jul–Aug 2003, Johns Hopkins Summer Workshop 2003
Member, Syntax for Statistical Machine Translation.
Along with a team of faculty members, graduate and undergraduate students, I explored statistical machine translation (MT) models that used various kinds of information from statistical parsers. The task was to find better translations within an n -best list produced by a baseline statistical MT system (we worked on the NIST 02/03 large data track for Chinese to English MT).
77. Jun–Jul 2002, Johns Hopkins Summer Workshop 2002
Member, Weakly Supervised Learning for Wide Coverage Parsing.
Along with Prof. Mark Steedman (Univ. of Edinburgh), I proposed a topic for inclusion in the 2002 workshop, and our proposal was funded as one of four out of twelve candidates. During the workshop, along with a team of faculty members, graduate and undergraduate students, I conducted experiments in improving accuracy of statistical parsing when faced with limited amounts of training data.
78. 1995–2002, The XTAG Project (<http://www.cis.upenn.edu/~xtag>)
Lead developer of an open-source parser for a wide-coverage linguistic grammar of English.
The XTAG project is an academic research project directed by Prof. Aravind Joshi. The project has built a wide-coverage lexicalized Tree Adjoining Grammar (TAG) and associated parser for English which handles naturally occurring text. It has been used for both recognition and generation. I designed a parsing algorithm for TAG and implemented a parser which showed significant speedup over the previous implementation.
79. Aug–Oct 1995, Penn MUC-6 Team
Team Member
Designed and implemented modules for the Univ. of Penn. Co-reference Resolution System that stood third at the Sixth Message Understanding Conference in the co-reference task.

80. 1991–1993, Centre for Development of Advanced Computing (CDAC)
University of Poona Campus, Pune 411 007, India
A research lab run by the Dept. of Electronics, Govt. of India.
Research Associate (Natural Language Group)
Implemented a morphological analyzer which extended Kimmo’s two-level model. Also worked on the implementation in Prolog of a tutoring system.

2.5 Invited Talks and Presentations

Invited Talks

81. *Combining Morpheme-based machine translation with post-processing morpheme prediction*
Language Technologies Research Centre and the Interactive Language Technologies Group of NRC-IIT.
Gatineau, Quebec. Mar 29, 2011.
82. *Translation beyond words: unsupervised word segmentation and statistical machine translation*
Department of Computer Science, University of Alberta. Nov 23, 2010.
83. *Latent TAG Derivations for Semantic Role Labeling*
University of Heidelberg, Heidelberg, Germany. Feb 4, 2010.
Cambridge University, Cambridge, U.K. May 12, 2010.
Open University, Milton Keynes, U.K. April 15, 2010.
84. *Bootstrapping a classifier using the Yarowsky algorithm*
University of Edinburgh, Edinburgh, Scotland. Oct 2, 2009.
85. *Lexicalized Tree-adjoining Grammar applied to semantic role labeling*
LORIA/INRIA. Nancy, France. Jun 5, 2008.
86. *Extensions of Regular Tree Grammars and their relation to Tree-adjoining Grammars*
Information Sciences Institute (ISI), Marina del Rey, USA. 16 Aug, 2007.
87. *Context-free languages is to Regular tree languages as Tree-adjoining languages is to what?*
University of Pennsylvania, Philadelphia, USA. Apr 21, 2007.
88. *Semi-supervised learning for statistical machine translation*
Machine Learning for Multilingual Information Access, NIPS 2006 workshop, Vancouver, BC. Dec 9, 2006.
89. *Bootstrapping statistical parsers from small data-sets*
University of Alberta, Artificial Intelligence Lab Seminar, Edmonton, AB. Mar 7, 2003.
90. *Combining Structural and Statistical Information: Relevance for Efficient Processing*
Seminar on Efficient Processing with High-Level Grammatical Formalisms, Schloss Dagstuhl, Germany. Oct 21, 1999.
91. *LR-parsing of very large Tree Adjoining Grammars*
DFKI, Saarbrücken, Germany. Aug 28, 1997.
92. *Synchronous Tree Adjoining Grammars and their application to Machine Translation*
Centre for Development of Advanced Computing, Pune, India. Jul 15, 1996.

Presentations

93. *Computational Constraints on Linguistic Descriptions*
Defining Cognitive Science at SFU series. Simon Fraser University, Burnaby, BC. Sep 26, 2007.
94. *A Statistical Parser for Hindi in ≤ 2 Weeks*
Joint work with P. Kanade, T. P. Reddy, M. Parakh, V. Mehta. Presentation at the Computational Linguistics Lunch (CLunch) Seminar at the University of Pennsylvania, 16 Jan 2002.

95. *A Statistical Parser for Hindi*
Joint work with P. Kanade, T. P. Reddy, M. Parakh, V. Mehta. Presentation of a 2 week project at the Corpus-Based NLP Workshop, Anna University, Chennai, India, 30 Dec 2001.
96. *Co-Training Methods for Statistical Parsing using Lexicalized Grammars*
Presentation at the NLP Open House 2000, IBM T. J. Watson Research Center, Hawthorne, 31 Oct 2000.
97. *Typing as a means for validating feature structures*
A. Sarkar and S. Wintner. Presentation by S. Wintner at CLIN99, Utrecht University, Netherlands, 10 Dec 1999.
98. *Structural Language Modeling using Stochastic Tree Adjoining Grammars*
Presentation at the AT&T Student Research Day 1999, 29 Oct 1999.
99. *Handling Coordination in Tree Adjoining Grammars*
Tutorial session at the Fourth Workshop on Tree Adjoining Grammars, TAG+ 4, Philadelphia, PA, 31 July 1998.
100. *Grammar Inference using TAGs and Conditions on their Consistency*
Poster Presentation at the First Annual Northeast Cognitive Science Society (NECSS) Graduate Conf. Cornell University, Ithaca, NY, 1-2 May 1998.

3 Service

3.1 Service to the Academic Community

Positions Held

- Secretary of the North American chapter of the Association for Computational Linguistics (NAACL), elected by the membership in 2009.

Grant Reviews

- MITACS College of Reviewers for the ACCELERATE program: April 2009 – April 2010
- Grant Reviewer, The Netherlands Organization for Scientific Research (NWO).
- Panel member and reviewer, NSF Career Review Panel (Human Language and Communication, Division of Information & Intelligent Systems) 2005.

Conference Chair

- Program Co-chair, CoNLL 2010 Conference on Natural Language Learning, Uppsala, Sweden, July 15-16, 2010.
- Faculty Advisor Program Co-chair, NAACL HLT 2009 Student Research Workshop, Boulder, Colorado, June 1-3, 2009.
- Program Co-chair, The 9th International Workshop on Tree Adjoining Grammar and Related Formalisms: TAG+9. University of Tübingen, June 2008.
- Area Chair, Parsing, Conf. on Empirical Methods in Natural Language Processing and the Conf. on Comp. Natural Language Learning: EMNLP-CoNLL 2007.
- Area Chair, Machine Learning Methods, Joint conference of the Intl. Committee on Comp. Ling. and the Assoc. for Comp. Ling.: ACL-COLING 2006.

Conference Organization

- Organizing Committee of the 1st CAIAC/Precarn AI Challenge, 2008-2009. (a Netflix style challenge to encourage AI interest in Canadian high school, undergraduate and graduate students).
- Local Preparation and Student Volunteer Coordinator, Human Language Tech. Conf. and Conf. on Empirical Methods in Natural Language Processing: HLT/EMNLP 2005.
- Local Organizer, The 7th International Workshop on Tree Adjoining Grammars and Related Formalisms: TAG+7, Vancouver, BC, 2004.
- Organizing Committee, 6th International Workshop on Tree Adjoining Grammars and Related Formalisms: TAG+6 2002, Venice, Italy.

Session Chair

- Annual meeting of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT). Boulder, Colorado. May 31-June 5, 2009.
- Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. EMNLP-CoNLL 2007, Prague, Czech Republic. August 2007.
- Joint Meeting of the Conference on Computational Linguistics and the Annual Meeting of the Association of Computational Linguistics: COLING-ACL 2006, Sydney, Australia. July 2006.
- Panel Member, Workshop on Effective Tools and Methodologies for Teaching NLP And Computational Linguistics.
- 7th International Workshop on Tree Adjoining Grammar and Related Formalisms, May 20-22, 2004. Simon Fraser University, Vancouver, Canada.

Program Committee

- Pacific Northwest Regional NLP Workshop 2010 (NW-NLP 2010). Redmond, Washington, April 23, 2010.
- 10th Conference on Tree Adjoining Grammars and Related Formalisms, TAG+ 10. Yale University, June 10-12, 2010.
- ACL 2010 Workshop on Applications of Tree Automata in Natural Language Processing. July 16, 2010. Uppsala, Sweden.
- MT Summit XII 2009. Ottawa, Ontario, Canada, August 26-30, 2009.
- NAACL HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing.
- Canadian Conference on Artificial Intelligence: AI'2007, '06, '05.
- Conference on Empirical Methods in Natural Language Processing: EMNLP 2009, 2004, '03, '02, '01.
- Conference on Natural Language Learning: CoNLL 2006, '05, '04.
- International Joint Conference on Natural Language Processing, IJCNLP 2007, '04.
- International Workshop on Tree Adjoining Grammar and Related Formalisms, TAG+8 2006.
- Workshop on Effective Tools and Methodologies for Teaching NLP And Computational Linguistics: TNLP-2005.
- FGVienna: 8th Conference on Formal Grammar, 2003.
- 6th Natural Language Processing Pacific Rim Symposium: NLPRS 2001.

Journal Reviews

- The Python Papers (TPP) and The Python Papers Source Codes (TPPSC) (2009)
- Linguistic Issues in Language Technology (LiLT) (2008)
- IEEE Transactions on Speech and Audio Processing (2007)
- IEEE Transactions on Knowledge and Data Engineering (2004)
- Computational Intelligence (2003)
- Natural Language Engineering (2003, 2009)
- Computational Linguistics (2001)

Book Reviews

- CSLI Lecture Notes 118 (Center for the Study of Language and Information, Stanford, CA). (2002)
- Statistical machine translation book, Cambridge University Press. (2007-2008)

Conference Reviews

- Annual Meeting of the Association for Computational Linguistics: ACL 2008, '07, '05, '04, '03, '02, '01, '00; ACL-COLING 1998.
- Annual Meeting of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2006, '04, '01, '00.
- International Conference on Computational Linguistics: COLING 2008, '00, '96.
- International Joint Conference on Artificial Intelligence: IJCAI 2007, '03.
- International Conference on Natural Language Processing (India): ICON 2007, '05, '04, '02
- Symposium on Latin American Theoretical INformatics: LATIN'2002.
- Fifteenth Annual Combinatorial Pattern Matching Symposium: CPM 2004.
- 7th International Workshop on Parsing Technologies: IWPT 2001.
- 6th Meeting on the Mathematics of Language (MOL6) 1998.

Professional Membership

- Secretary of the North American chapter of the Association for Computational Linguistics, elected in 2009.
- Member, Association for Computational Linguistics, since 1993.
- Member, Neural Information Processing Systems Foundation, since 2002.

January 16, 2012