**Harvard Medical School Curriculum Vitae**

**Date Prepared:** November 7, 2014

**Name:** Tim Miller

**Office Address:** 300 Longwood Avenue, Boston, MA 02115

**Home Address:** 9 Moraine St., Boston, MA 02130

**Work Phone:** 617-919-1223

**Work E-Mail:** timothy.miller@childrens.harvard.edu

**Work FAX:**

**Place of Birth:** Green Bay, WI

**Education**

| Year | Degree (Honors) | Field of Study (Thesis advisor for doctoral research degrees) | Institution |
|---|---|---|---|
| 2003 | BS (Magna Cum Laude) | Computer Science | Marquette University |
| 2007 | MS | Computer Science (William Schuler) | University of Minnesota |
| 2010 | PhD | Computer Science (William Schuler) | University of Minnesota |

**Appointments at Hospitals/Affiliated Institutions**

| | | | |
|---|---|---|---|
| 2011 | Research Fellow | Informatics Program | Boston Children's Hospital |
| 2011 | Research Fellow | Harvard Medical School | Boston, MA |
| 2014 | Instructor | Informatics Program | Boston Children's Hospital |
| 2014 | Instructor | Harvard Medical School | Boston, MA |

**Other Professional Positions**

| Year(s) | Position Title | Institution |
|---|---|---|
| 2010-2011 | Associate Scientist | University of Wisconsin - Milwaukee |

**Editorial Activities**

2011 Reviewer 2011 i2b2/VA Challenge on Coreference Resolution
2011 Reviewer Biomedical NLP Workshop at Recent Advances in Natural Language Processing Conference (RANLP)
2011 Reviewer Empirical Methods in Natural Language Processing Conference (EMNLP)
2012 Reviewer IEEE Conference on Healthcare Informatics, Imaging, and Systems Biology
2012 Reviewer American Medical Informatics Association Annual Symposium
2012 Program Committee Computational Semantics for Clinical Text Workshop
2013 Reviewer Biomedical NLP Workshop at Recent Advances in Natural Language Processing
2013 Reviewer American Medical Informatics Association Annual Symposium

2014 Reviewer Journal of the American Medical Informatics Association (JAMIA)
2014 Reviewer American Medical Informatics Association Annual Symposium
2014 Reviewer Journal of Biomedical Informatics (JBI)

**Honors and Prizes**

| Year | Name of Honor/Prize | Awarding Organization | Achievement for which awarded |
|---|---|---|---|
| 2004-2006 | IGERT Training Fellowship | National Science Foundation | Research in Computational Neuroscience |

**Report of Funded and Unfunded Projects**

**Funding Information**
**Current:**
1U24CA184407-01 Crowley, Savova (MPI) 06/01/2014 - 05/31/2019 2.4 calendar months

NCI, NIH Annual Direct: $415,000 Total Direct: $2,856,017
Cancer Deep Phenotype Extraction from Electronic Medical Records

Precise phenotype information is needed to advance translational cancer research, particularly to unravel the effects of genetic, epigenetic, and systems changes on tumor behavior and responsiveness. Examples of phenotypic variables in cancer include: tumor morphology (e.g. histopathologic diagnosis), co-morbid conditions (e.g. associated immune disease), laboratory findings (e.g. gene amplification status), specific tumor behaviors (e.g. metastasis) and response to treatment (e.g. effect of a chemotherapeutic agent on tumor). Current models for correlating EMR data with –omics data largely ignore the clinical text, which remains one of the most important sources of phenotype information for cancer patients. Unlocking the value of clinical text has the potential to enable new insights about cancer initiation, progression, metastasis, and response to treatment. We propose further collaboration of two mature informatics groups with long histories of developing open-source natural language processing (NLP) software (Apache cTAKES, caTIES and ODIE) to extend existing software with new methods for cancer deep phenotyping.

1 R01 GM103859-01A1 (Denny/Xu/Pathak) 09/18/2014 – 08/31/2018

NIH/NIGMS Annual direct: $75,218 Total Direct: $7,127,855

Informatics Tools for Pharmacogenomic Discovery using Practice-based Data

Goal: The Informatics for Integrating Biology and the Bedside (i2b2), a National Center for Biomedical Computing based at Partners Healthcare System, has developed a scalable informatics framework to enable clinical researchers to use existing EMR data for genomic knowledge discovery of diseases. In this study, we will collaborate with i2b2 to extend its informatics framework to the pharmacogenomics domain, by developing new natural language processing, ontology components, and user-friendly interfaces, and then apply these tools to real-world pharmacogenomic studies.

**Past:**
1R01LM010090-01 Savova, Palmer (PI) 07/01/10-09/29/14

NIH/NLM Direct: $584,450

Temporal Relation Discovery for Clinical Text

The goal of our current proposal is to automatically discover temporal relations from clinical free text and create a timeline. Temporal relations are of prime importance in biomedicine as they are intrinsically linked to diseases, signs and symptoms, and treatments. Understanding the timeline of clinically relevant events is key to the next generation of translational research where the importance of generalizing over large amounts of data holds the promise of deciphering biomedical puzzles. The project is a collaborative effort between Mayo Clinic and University of Colorado (Profs. Martha Palmer, James Martin and Wayne Ward).

2010-2014 ONC $141,103/yr
90TR0002/01
SHARP Area 4: Secondary use of the EMR (Chute)
This project focuses on building a framework of open-source services that can be dynamically configured to transform EHR data into standards-conforming, comparable information suitable for large-scale analyses, inferencing, and integration of disparate health data. The clinical narrative and NLP methods for its processing are a central piece towards data normalization.

2011-2014 U54LM008748 National Library of Medicine $357,737
Informatics for Integrating Biology and the Bedside
The goal of this consortium grant is to advance clinical Research in the genomic era. The NLP component focuses on a portable, extensible and modular framework for processing the clinical narrative and extracting a variety of key information from it. The toolset will be released as part of the i2b2 open source framework.

**Submitted:**

Review pending — Unsupervised induction of clinical grammars for next generation information extraction

National Institutes of Health/National Library of Medicine
PI – Direct costs requested: $1,260,000
This grant proposes machine learning methods for learning language structure that can be applied to new domains without creating hand-labeled annotations for each domain. These methods can improve accuracy and speed of adapting NLP methods in new domains, in turn broadening the potential impact of NLP methods on clinical research.

## Report of Local Teaching and Training

### Teaching of Students in Courses

Year(s) Course title
Type of student/audience Role in course Level of effort

**University of Minnesota**

| | |
|---|---|
| 2007 | Computer Science 4041-Algorithms and Data Structures |

| | | |
|---|---|---|
| Upper-level undergraduates | Teaching Assistant | 50% |

| | | |
|---|---|---|
| 2007 | Computer Science 5381-Computational Techniques for Genomics | |
| Graduate students | Teaching Assistant | 25% |

| | | |
|---|---|---|
| 2007 | Computer Science 5541-Natural Language Processing | |
| Graduate students | Teaching Assistant | 25% |

| | | |
|---|---|---|
| 2009-2010 | Computer Science 1902-Structure of Computer Programming II | |
| Lower-level undergraduates | Teaching Assistant | 50% |

## Report of Regional, National and International Invited Teaching and Presentations

No presentations below were sponsored by outside entities

**National:**

| | |
|---|---|
| 2012 | Tutorial: Active Learning for Clinical Coreference Resolution. |
| | UC-San Diego, (iDASH - NLP Annotation Workshop) |

| | |
|---|---|
| 2012 | Coreference Resolution in cTAKES |
| | Boston, MA (i2b2 Academic Users Group Conference) |

**Regional:**

| | |
|---|---|
| 2012 | Clinical NLP and Machine Learning |
| | Cambridge, MA (Harvard Machine Learning Tea) |

## Report of Scholarship

**Publications**

William Schuler & **Tim Miller**. Integrating Denotational Meaning into a DBN Language Model. Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech/Interspeech'05), Lisbon, Portugal, 2005.

Shana Watters, **Tim Miller**, Praveen Balachandran, William Schuler & Richard Voyles. Exploiting a Sensed Environment to Improve Human-Agent Communication. Proceedings of the 4th Interna- tional Joint Conference on Autonomous Agents and Multi-Agent Systems (AA/MAS'05), Utrecht, Netherlands, 2005.

William Schuler, **Tim Miller**, Andy Exley & Steven Wu. Dynamic Evidence Models in a DBN Phone Recognizer. Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP/Interspeech'06), Pittsburgh, PA, 2006.

**Tim Miller**, Lane Schwartz & William Schuler. Incremental Semantic Models for Continuous Context-

Sensitive Speech Recognition. SRSL07 - Workshop on the Semantic Representation of Spoken Language, Salamanca, Spain, 2007.

**Tim Miller**, William Schuler & Andy Exley. Elements of a Spoken Language Programming Inter- face for Robots. Proceedings of HRI 2007, Washington, D.C., 2007.

**Tim Miller** & William Schuler. A Syntactic Time-Series Model for Parsing Fluent and Disfluent Speech. Proceedings of COLING, Manchester, UK, 2008. (26.8% acceptance rate)

William Schuler, Samir AbdelRahman, **Tim Miller** & Lane Schwartz. Toward a Psycholinguistically-Motivated Model of Language Processing. Proceedings of COLING, Manchester, UK, 2008. (26.8% acceptance rate)

**Tim Miller** & William Schuler. A Unified Model for Parsing Fluent and Disfluent Speech. 46th Annual Meeting of the Association for Computational Linguistics (ACL '08), Columbus, Ohio, USA, 2008. (25% acceptance rate)

**Tim Miller**. Word Buffering Models for Improved Speech Repair Parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 2009. (20% acceptance rate for oral presentations)

**Tim Miller**, Luan Nguyen & William Schuler. Parsing Speech Repair without Specialized Gram- mar Symbols. 47th Annual Meeting of the Association for Computational Linguistics (ACL '09), Singapore, 2009. (25% acceptance rate)

**Tim Miller**. Improved Syntactic Models for Parsing Speech with Repairs. Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL), Boulder, CO, 2009. (28.8% acceptance rate)

William Schuler, Samir AbdelRahman, **Tim Miller** & Lane Schwartz. Broad-Coverage Incremental Parsing using Human-Like Memory Constraints. Computational Linguistics, 2010.

**Tim Miller**, James J. Cimino, Kourosh Ravvaz, and Hong Yu. "An Investigation into the Feasibility of Spoken Clinical Question Answering." In *AMIA Annual Symposium Proceedings*, vol. 2011, p. 954. American Medical Informatics Association, 2011.

Chen Lin, Helena Canhao, **Timothy Miller**, Dmitriy Dligach, Robert M. Plenge, Elizabeth W. Karlson, Guergana Savova. 2012. Feature Engineering and Selection for Rheumatoid Arthritis Disease Activity Classification Using Electronic Medical Records. International Conference of Machine Learning (ICML) Workshop on Machine Learning for Clinical Data.

Chen Lin, **Timothy Miller**, Dmitriy Dligach, Robert M. Plenge, Elizabeth W. Karlson, Guergana Savova. 2012. Maximal Information Coefficient for Feature Selection for Clinical Document Classification. International Conference on Machine Learning Workshop on Machine Learning for Clinical Data.

**Timothy A. Miller**, Dmitriy Dligach, and Guergana K. Savova. "Active learning for coreference resolution." In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 73-81. Association for Computational Linguistics, 2012.

Jiaping Zheng, Wendy W. Chapman, **Timothy A. Miller**, Chen Lin, Rebecca S. Crowley, and Guergana K. Savova. "A system for coreference resolution for the clinical narrative." *Journal of the American Medical Informatics Association* 19, no. 4 (2012): 660-667.

Balaji Polepalli Ramesh, Rashmi Prasad, **Tim Miller**, Brian Harrington, and Hong Yu. "Automatic discourse connective detection in biomedical text." *Journal of the American Medical Informatics Association* 19, no. 5 (2012): 800-808.

Dmitriy Dligach, Steven Bethard, Lee Becker, **Timothy A. Miller**, Guergana K. Savova. Discovering Body Site and Severity Modifiers in Clinical Text. *Journal of the American Informatics Association*, 21(3):448-454, 2013.

Chen Lin, Elizabeth K. Karlson, Helena Canhao, **Timothy A. Miller**, Dmitriy Dligach, Pei Jun Chen, Raul Natanael Guzman Perez, Tianxi Cai, Michael E. Weinblatt, Nancy A. Shadick, Robert M. Plenge, Guergana K. Savova, Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records, Plos One. 2013 (in press).

**Timothy A. Miller**, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana K. Savova. "Discovering Narrative Containers in Clinical Text." In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, 2013.

Pascal B Pfiffner, JiWon Oh, **Timothy A. Miller**, Kenneth D. Mandl. ClinicalTrials.gov as a Data Source for Semi-Automated Point-Of-Care Trial Eligibility Screening. *PLoS One*, 9(10). 2014.

Chen Lin, **Timothy Miller**, Alvin Kho, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, and Guergana Savova. Descending-Path Convolution Kernel for Syntactic Structures. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 81-86, 2014.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, **Timothy Miller**, Chen Lin, Guergana Savova, James Pustejovsky. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143-154, 2014.

Stephen Wu, **Timothy Miller**, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, Cheryl Clark. Negation's Not Solved: Generalizability versus Optimizablity in Clinical Natural Language Processing. *PLoS One*. In press. 2014.

Chen Lin, Elizabeth W. Karlson, Dmitriy Dligach, Monica P. Ramirez, **Timothy A. Miller**, Huan Mo, Natalie S. Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, Guergana K. Savova. Automatic Identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Informatics Association.* In press. 2014.

**Thesis**

Generative Models of Disfluency. University of Minnesota, May 2010.

**Abstracts**

**Timothy Miller**, Dmitriy Dligach, Steven Bethard, Sameer Pradhan, Chen Lin, and Guergana K. Savova. Discovering Time Expressions in Clinical Text. Annual Symposium of the American Medical Informatics Association. 2013.

Stephen Wu, **Timothy Miller**, James Masanz, Matthew Coarr, David Carrell, Scott Halgrim, David Harris, and Cheryl Clark. Negation's Not Solved: Reconsidering negation annotation and evaluation. Annual Symposium of the American Medical Informatics Association. 2013.

Pascal B Pfiffner, JiWon Oh, **Timothy A Miller**, and Kenneth D Mandl. ClinicalTrials.gov as a data source for point-of-care trial recruitment. ASCI/AAP Joint Meeting, Chicago, 2014.

## Narrative Report

I have now been working in the field of biomedical natural language processing for three years, with a focus on clinical NLP. Coming from graduate work in computer science and general domain NLP, I have learned an immense amount in my transition to research in the clinical domain. I have been fortunate to have an excellent mentor in Dr. Guergana Savova in addition to many excellent collaborators and colleagues from the Harvard Medical School community and around the world.

My plan has been to learn about the applications and intricacies in the clinical domain while also broadening my knowledge of the state of the art in the general domain, so that I can port cutting edge techniques from the general domain while adapting and extending where necessary for the clinical domain. I have been successful in that endeavor, publishing on a variety of problems and expanding my repertoire of technical expertise. My most recent published work (Miller et al., BioNLP 2013) extended work from the general domain on the use of tree kernels (a method for incorporating syntactic features in machine learning) to the important clinical NLP problem of relating events to times in clinical narratives. Additional work on extracting temporal information from clinical text for that project (THYME) is underway and nearly ready to publish. I have also worked on the coreference resolution problem, adapting general domain machine learning systems with clinically relevant features, including novel features like semantic similarity based on Wikipedia articles. In addition to these main foci I have collaborated closely with fellow lab members on the tasks of phenotype extraction, relation extraction, and negation detection.

One of my guiding interests going forward is the optimal use of annotated resources and making use of unlabeled data. To that end I have worked on Active Learning, a method for optimally selecting training instances for machine learning systems. This has led to a publication related to coreference resolution, and current work applies this framework to the problem of patient phenotyping, while also taking advantage of other unlabeled resources. Going forward I am also pursuing methods that mine web data to extract information that is difficult or tedious to annotate. Human annotators are indispensable for clinical NLP, but for some tasks it is sufficient or even preferable to utilize noisy unlabeled data in massive quantities.