OMB No. 0925-0001/0002 (Rev. 08/12 Approved Through 8/31/2015)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.  
Follow this format for each person. DO NOT EXCEED FIVE PAGES.

NAME: Dmitriy Dligach

eRA COMMONS USER NAME (credential, e.g., agency login): dmitriydligach

POSITION TITLE: Assistant Professor, Computer Science, Loyola University Chicago, Chicago, IL  
 Assistant Professor, Public Health Sciences, Loyola University Medical School, Chicago, IL

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)

| INSTITUTION AND LOCATION | DEGREE  (if applicable) | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| Loyola University Chicago | B.S. | 01/99 | Computer Science |
| State University of New York at Buffalo | M.S. | 12/04 | Computer Science |
| University of Colorado Boulder | Ph.D. | 09/10 | Computer Science |

# A. Personal Statement

# My graduate and post-graduate research spans machine learning, natural language processing (NLP), and medical informatics. My thesis work focused on developing new methods for resolution of semantic ambiguities. For my Ph.D. dissertation, I developed and implemented new approaches for representing textual data leading to highly accurate and robust machine learning models. As a researcher at Boston Children’s Hospital and Harvard Medical School I worked on applications of machine learning and NLP to health and medicine. Specifically, I was involved in several projects that focused on deep semantic analysis of clinical texts including relation extraction between medical entities, phenotyping, and other types of data mining. I lead the development of the methods and software for relation extraction; our best performing methods were implemented and released open-source as a part of the clinical Text Analysis and Knowledge Extraction System (cTAKES) – the most widely adopted software for clinical text processing. The relation extractor later became the basis for the temporal relation extraction software developed for the THYME (Temporal Histories of Your Medical Event) project. In addition to my extensive research expertise in NLP and machine learning, I bring years of software development experience that is crucial for creating well-designed, reliable, and interoperable software that we plan to deliver for this project. In January, 2016 joined the computer science department at Loyola University Chicago as an assistant professor, where I continue my research in medical informatics, natural language processing, and machine learning. I currently also have a secondary appointment at Loyola Stritch School of Medicine. My expertise in semi-supervised methods and clinical information extraction (as indicated by my publication record) build a solid foundation for my contribution to the specific aim 2 of this proposal. Dr. Miller (PI) and I have an extensive history of collaboration that began in 2011 when I joined the NLP lab at Boston Children’s Hospital (BCH) and Harvard Medical School. Our time at BCH had been extremely productive and resulted in a number of joint publications. In the five years that I have known Dr. Miller, I have witnessed on numerous occasions his remarkable leadership abilities: Dr. Miller led a team of researchers who under his guidance built several state-of-the art NLP systems such as the cTAKES coreference system, the negation and uncertainty system, and several components of the THYME temporal information extraction system. In addition, Dr. Miller led the development of several open source projects such as ClearTK and libsvm-java-kernel. His efficient leadership and massive technical competence leave no doubt in my mind that the time has come for him to transition to the role of a PI of a major NIH-funded grant.

# B. Positions and Honors

**Positions and Employment**

1999-2000 Development Analyst, Commonwealth Edison, Chicago, IL

2000-2003 Senior Software Engineer, Motorola, Inc., Arlington Heights, IL

2005 Research Scientist, Semantic Discovery, Boulder, CO

2005-2010 Research Assistant, University of Colorado, Boulder, CO

2010-2011 Postdoctoral Research Associate, University of Colorado, Boulder, CO.

2011-2014 Research Fellow, Boston Children’s Hospital and Harvard Medical School, Boston, MA

2014-2015 Associate Scientific Researcher, Boston Children’s Hospital, Boston, MA

2014-2015 Instructor, Harvard Medical School, Boston, MA

2016- Assistant Professor, Loyola University, Chicago, IL

**Select Other Experience and Professional Memberships**

2015 American Medical Informatics Association (AMIA) Symposium. Paper Reviewer.

2015 Association for Computational Linguistics (ACL) conference. Program Committee.

2015 Association for Computational Linguistics (ACL) conference. Paper Reviewer.

2015 SemEval 2015. Paper Reviewer.

2015 International Conference on Computational Semantics (IWCS). Program Committee.

2015 International Conference on Computational Semantics (IWCS). Paper Reviewer.

2013 RANLP NLP for Medicine and Biology Workshop. Paper Reviewer.

2013 American Medical Informatics Association (AMIA) Symposium. Paper Reviewer.

2013 International Conference on Computational Semantics (IWCS). Program Committee.

2013 International Conference on Computational Semantics (IWCS). Paper Reviewer.

2011 i2b2/VA/Cincinnati 2011 Shared Task and Workshop. Paper Reviewer.

2011 Computational Linguistics. Paper Reviewer.

2010 SemEval 2010. Program Committee.

2010 SemEval 2010. Paper Reviewer.

2010 SemEval 2010 Tasks 14: Word Sense Induction & Disambiguation Task. Organizing Committee.

# C. Contribution to Science

**1. Efficient and cost-effective machine learning model development**: Supervised learning is widely used in NLP. However, it requires annotated training data, which is expensive to obtain. How does one choose the data for annotation? Random sampling is a common approach but not the most efficient one. Various types of selective sampling can be used to achieve the same level of performance as random sampling but with less data. Active learning is one type of selective sampling, but in many situations it is not practical (e.g. in a multiple-annotator, double-annotation environment). I experimented with various more practical alternatives to the classic active learning, and showed that it can lead to a sizable reduction in the amount of annotation both in the newswire domain [2] and in the clinical domain [3]. I also evaluated the relative benefits of various popular annotation practices such as single annotation and double annotation and ways of reducing their costs. I proposed several approaches to reducing the amount of the second round of annotation. By double annotating only a carefully selected subset of potentially erroneous and hard-to-annotate single annotated examples, the amount of the second round of annotation can be reduced by more than half without sacrificing the performance [4]. More recently, I conducted a series of experiments with semi-supervised learning and demonstrated that unlabeled data can help improve the performance of phenotyping models while significantly cutting the cost of annotation [1].

1 **Dligach D**, Miller TA, Savova GK. Semi-supervised Learning for Phenotyping Tasks. In: *Proceedings of*

*American Medical Informatics Association Symposium 2015*. San Francisco, CA. November, 2015.

2 **Dligach D**, Palmer M. Good seed makes a good crop: accelerating active learning using language modeling. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. 2011. 6–10.

3 **Dligach D**, Miller TA, Savova GK. Active Learning for Phenotyping Tasks. In: *Proceedings of the Workshop on NLP for Medicine and Biology associated with RANLP*. 2013. 1–8.

4 **Dligach D**, Palmer M. Reducing the need for double annotation. In: *Proceedings of the 5th Linguistic Annotation Workshop*. 2011. 65–73.

**2. Deep semantic analysis of clinical narratives:** Current state-of-the-art clinical natural language processing systems are capable of extracting medically relevant textual entities from electronic health records, such as “severe pain” and “left arm”. My goal as a researcher is to develop approaches that make a deeper level of semantic analysis possible. In the examples mentioned above, I would like to be able to discover that the “pain” is located in the right arm and that the “pain” is “severe”. Toward that end, I led the development of the methodology for determining the relations between clinical entities [1]. The proposed methodology was implemented in software as a part of Apache cTAKES (<http://ctakes.apache.org>) and released open-source, creating an off-the-shelf solution fine-grained text mining.

5 **Dligach D**, Bethard S, Becker L, *et al.* Discovering body site and severity modifiers in clinical texts. *J Am Med Inform Assoc* 2014;**21**:448–54.

**3.** **Temporal analysis of clinical narratives**: The notion of temporality has recently received much attention in clinical informatics community due to the fact that many important phenotypes are time-sensitive. For example, discovering patients that had adverse drug reactions requires analyzing the text to determine whether the patient had been exposed to the drug *before* (and not after) the adverse effects were detected. I worked on incorporating temporal information extracted from the text into the model identifying the patients with drug-induced liver toxicity [2]. I also focused on developing new methods for extracting temporal information from text [1,3,4].

1 \*Lin C, \***Dligach D**, \*Miller TA, \*Bethard S, Savova GK. Multilayered Temporal Modeling for the Clinical

Domain. *Journal of American Medical Informatics Association*. 2015. doi:

http://dx.doi.org/10.1093/jamia/ocv113. \*Indicates co-first authors

2 \*Lin C, \*Karlson EW, \***Dligach D**, *et al.* Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *J Am Med Informatics Assoc* 2014;:amiajnl – 2014. \*Indicates co-first authors

3 Miller TA, Bethard S, **Dligach D**, *et al.* Discovering narrative containers in clinical text. *ACL 2013* 2013;:18.

4 Lin C, Miller T, Kho A, Bethard S, **Dligach D**, *et al.* Descending-Path Convolution Kernel for Syntactic Structures. *ACL* 2014;**1**:81–6.

**4. Resolution of lexical ambiguities:** Resolution of lexical ambiguities has for a long time been viewed as a key problem in NLP that tests our ability to capture and represent semantic knowledge and learn from linguistic data. I proposed and evaluated a novel type of semantic feature that can be extracted from unlabeled corpora. These features, known as Dynamic Dependency Neighbors (DDNs), represent the meaning of nouns by capturing co-occurrence statistics of verbs and their noun objects. One important characteristic of the DDNs is that they capture lexical semantic knowledge without relying on hand-crafted linguistic resources, the traditional source of this type of knowledge. This makes the DDNs potentially applicable to many resource-poor languages and domains. I showed that incorporating the DDNs into the classification framework leads to performance improvements both in the context of word sense disambiguation [1] and VerbNet class disambiguation [2,3].

1 **Dligach D**, Palmer M. Novel semantic features for verb sense disambiguation. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. 2008. 29–32.

2 Brown SW, **Dligach D**, Palmer M. VerbNet Class Assignment As a WSD Task. In: *Proceedings of the Ninth International Conference on Computational Semantics*. Stroudsburg, PA, USA: : Association for Computational Linguistics 2011. 85–94.http://dl.acm.org/citation.cfm?id=2002669.2002679

3 Brown SW, **Dligach D**, Palmer M. VerbNet class assignment as a WSD task. In: *Computing Meaning*. Springer 2014. 203–16.

**Complete List of Published Work in Google Scholar:**

<https://scholar.google.com/citations?user=mkLwEPkAAAAJ&hl=en>

**D. Research Support**

**Completed Research Support**

97324-01 Dligach, Miller (PI) 05/2015-12/2015

Boston Children’s Hospital, The Program for Patient Safety and Quality

**Natural Language Processing for Quality Metrics**

The quality of patient care in the United States remains substandard: according to a New England Journal of Medicine study, patients receive proper care only in about 55% of the cases. Improving the quality of care is therefore an urgent goal of the health care system. Computing quality measures is however a complicated process, which in many cases involves a manual examination of a large number of electronic health records (EHRs). The goal of this project is to evaluate the use of automatic natural language processing (NLP) and machine learning methods for computing quality measures.

1U24CA184407-01 Crowley, Savova (MPI) 06/2014 - 05/2019

NCI, NIH

**Cancer Deep Phenotype Extraction from Electronic Medical Records (DeepPhe)**

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.

1R01LM010090-01 Savova, Palmer (PI) 07/2010-09/2014

National Institutes of Health/National Library of Medicine

**Temporal Relation Discovery for Clinical Text (THYME)**

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).

Role: Co Investigator

U54LM008748 Kohane (PI) 09/2010-09/2014

National Library of Medicine

**Informatics for Integrating Biology and the Bedside (i2b2)**

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.

Role: Co Investigator

R01GM090187 Chapman, Elhadad, Savova (PI) 7/2010–6/2014

National Institutes of Health   
**Annotation, development, and evaluation for clinical information extraction (ShARe)**Over the last two decades, several research groups have developed NLP tools for clinical notes, but a major bottleneck preventing progress in clinical NLP is the lack of standard, annotated data sets for training and evaluating NLP applications. Without these standards, individual NLP applications abound without the ability to train different algorithms on standard annotations, share and integrate NLP modules, or compare performance. We propose to develop standards and infrastructure that can enable technology to extract scientific information from textual medical records, and we propose the research as a collaborative effort involving NLP experts across the U.S.

Role: Co Investigator

90TR0002 (PI: Chute) 4/01/2010 – 9/30/2013

ONC

**SHARP Area 4: Secondary Use of EHR Data**

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.

Role: Postdoctoral Research Fellow

1 RC1 LM010608 (PI: Martin) 9/30/2009 – 9/29/2011

NIH

**Multi-Source Integrated Platform for Answering Clinical Questions**

The goal of this research is to develop a question answering system builds on existing mature technologies. Role: Postdoctoral Research Fellow

HR0011-06-C-0022 (PI: Palmer) 10/01/2006 – 4/30/2011

DARPA

**Autonomous Global Integrated Language Exploitation**

The goal of this project was sense tagging and PropBanking for English, Chinese and Arabic.

Role: Graduate Research Assistant

IIS 0715078 (PI: Palmer) 9/01/2004 – 11/30/2009

National Science Foundation

**Advancing the Performance of WSD Systems through More Consistent Criteria for Sense Distinctions**

The goal of this project is the annotation of replicable sense tags as measured through ITA and the training of accurate automatic sense taggers.

Role: Graduate Research Assistant