OMB No. 0925-0001 and 0925-0002 (Rev. 10/15 Approved Through 10/31/2018)

BIOGRAPHICAL SKETCH

**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   
 Assistant Professor, Public Health Sciences, Loyola University School of Medicine

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 artificial intelligence (AI), machine learning, natural language processing (NLP), and medical informatics. 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 Events) project. In addition to my extensive research expertise in NLP and machine learning, I bring years of software development experience that is crucial for building well-designed, reliable, and interoperable software that we plan to create for this project.

In January, 2016 I joined the computer science department at Loyola University Chicago as an assistant professor, where I continue my research in medical informatics, NLP, and machine learning. I currently also have a secondary appointment at Loyola Stritch School of Medicine and Boston Children’s Hospital. Recently, I have conducted experiments focusing on applications of deep neural networks ("deep learning") to clinical NLP. These experiments demonstrated that neural models perform approximately as well as the state-of-the-art and in many cases show improvements over the state-of-the-art, while completely eliminating the need for manual feature engineering. A number of these works are now published in the proceedings of top NLP, AI, and medical informatics conferences. I recently also experimented with learning patient representations from the text of EHRs. I used these patient representations in the context of a standard phenotyping task with a strongly positive result. This work was just presented at a major NLP conference. My expertise in machine learning and clinical information extraction build a solid foundation for my contribution to this proposal.

I have been part of the THYME project since the inception and while working as a researcher at Boston Children's Hospital and Harvard Medical School. Since joining Loyola University, I have continued my collaboration with all team members of the THYME project. We have developed a number of novel methods for temporal information extraction and other clinical NLP tasks and have a strong publication record in these areas. I am looking forward to continuing this collaboration to help develop a new generation of methods for temporal information extraction.

# 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

2015- Secondary Appointment, Boston Children’s Hospital, Boston, MA

2016- Assistant Professor, Loyola University School of Medicine, Chicago, IL

2016- Assistant Professor, Loyola University, Chicago, IL

**Select Other Experience and Professional Memberships**

2018 International Conference on Computational Linguistics (COLING) conference. Paper Reviewer.

2018 Association for Computational Linguistics (ACL) conference. Program Committee.

2018 Association for Computational Linguistics (ACL) conference. Paper Reviewer.

2017 Association for Computational Linguistics (ACL) conference. Program Committee.

2017 Association for Computational Linguistics (ACL) conference. Paper Reviewer.

2016 Clinical NLP Workshop at COLING. Paper Reviewer.

2016 Clinical NLP Workshop at COLING. Program Committee.

2016 Association for Computational Linguistics (ACL) conference. Paper Reviewer.

2016 American Medical Informatics Association (AMIA) Symposium. Paper Reviewer.

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 [3] and in the clinical domain [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 [2]. In my most recent work [1], I explore the use of deep neural networks for learning patient representations and for transfer learning. Mining electronic health records for patients who satisfy a set of predefined criteria is known in medical informatics as phenotyping. Phenotyping has numerous applications such as outcome prediction, clinical trial recruitment, and retrospective studies. Supervised machine learning for phenotyping typically relies on sparse patient representations such as bag-of-words. I considered an alternative that involved learning patient representations. I developed a neural network model for learning patient representations and showed that the learned representations were general enough to obtain state-of-the-art performance on a standard comorbidity detection task.

1 **Dligach D**, Miller TA. Learning Patient Representations from Text. In: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*. 2018. New Orleans, LA. June, 2018.

2 **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.

3 **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.

4 **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.

**2. Information extraction from clinical texts:** 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 [1,2]. 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.

1 **Dligach D**, Miller T, Lin C, et al. Neural temporal information extraction. *Proceedings of the 15th Annual Meeting of the European Association for Computational Linguistics.* Valenica, Spain. 2017.

2 **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 Viani N, Miller T, **Dligach D**, et al. Recurrent Neural Network Architectures for Event Extraction from Italian Medical Reports. In *AIME 2017: Proceedings of the 16th Conference on Artificial Intelligence in Medicine*. Vienna, Austria. June, 2017.

**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 [1,3,4,5].

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 Lin C, Miller TA, **Dligach D**, Bethard S, Savova GK. Representations of Time Expressions for Temporal Relation Extraction with Convolutional Neural Networks. In Proceedings of the 2017 Workshop on Biomedical Natural Language Processing (BioNLP 2017) held in conjunction with ACL-2017. August 2017. Vancouver, Canada.

4 Lin C, Miller TA, **Dligach D**, Bethard Steven, Savova GK. Improving Temporal Relation Extraction with

Training Instance Augmentation. In: *Proceedings of the 2016 Workshop on Biomedical Natural Language*

*Processing (BioNLP 2015) held in conjunction with ACL-2016*. August 2016. Berlin, Germany.

**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.

**D. Research Support**

**Completed Research Support**

97324-01 PI: Dligach, Miller 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.

2R01LM010090-05 PI: Savova 09/30/15 – 09/29/18

NLM, NIH

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

Renewal: The goal of this 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. Methods are machine learning.

Role: Co-Investigator

5U24CA184407-02 Crowley, Savova (MPI) 05/01/2014 – 04/30/2019

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 Current models for correlating EMR data with –omics data largely ignore the clinical text. Unlocking the value of clinical text has the potential to enable new insights about cancer initiation, progression, metastasis, and response to treatment. We propose to extend existing software with new methods for cancer deep phenotyping.

Role: Co-Investigator

1 R01 LM010090 PI: Savova, Palmer 7/01/2010 – 9/29/2014

HHS

**Temporal Relation Discovery for Clinical Text**

The goal of this 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. Methods are machine learning.

Role: Postdoctoral Research Fellow

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: Postdoctoral Research Fellow

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