Northeastern University
College of Engineering

Multidisciplinary Graduate Engineering
Course Syllabus

Course Information

Natural Language Processing

Course Number

INFO7610

Credit Hours

4

Instructor Information

Dino Konstantopoulos
Dino.k@northeastern.edu

Technical/Course Materials Requirements

The class requires an installation of Anaconda Python and a laptop with an i5 or better/equivalent processor, Mac or Windows. Students require a background in data science and/or Machine Learning, preferably with classes in Algorithms and data structures. These requirements may be waved at the discretion of the department, with the professor’s recommendation.

Course Description/Prerequisite

We all know Machine Learning as a stack of either dense model, convolutional model (CNN), or recurrent model (RNN) networks. In fact, vision ML is often all about CNNs, text processing often about RNNs, and the rest about dense (fully connected) networks. However, new paradigms have emerged in Natural Language Processing (NLP), with fully connected networks called ***transformers***.

In late 2017, a paper entitled "Attention Is All You Need" took the deep learning community by storm and is now the go-to method for sequence transduction (NLP) tasks, like translation, speech-to-text, and chatbots. Transformer models leverage dense networks with a cleverly-designed attention mechanism. The core idea behind the Transformer model is "self-attention": The ability to attend to different positions of the input sequence to compute a semantically correct representation of that sequence. It’s a bit like looking at a picture of a scene replete of activity and paying attention to details all over the picture in order to understand what is really going on.

Transformer models do not use RNNs nor CNNs. And yet they yield higher translation quality, require less computation to train, and are a much better fit for modern machine learning hardware, speeding up training by up to an order of magnitude. It powers google translate.

In this advanced deep learning class, we will study the 3 revolutions in NLP that made Transformers possible: Beginning with sequence to sequence RNNs, followed by word embeddings that capture semantics from syntax, and culminating with self-attention and Transformer models like Google's BERT and OpenAI's GPT-3. We'll study the foundations of the Transformer architecture by coding implementations in class using both tensorflow and pytorch, the two leading ML frameworks. We'll translate Chinese and Hindi to English, understand how Alexa recognizes speech and code our very own Alexa, generate fake news from NLP corpora, and train chatbots. We'll also uncover that a neural model that is trained on Natural Language is a better neural model for almost *any* task, which suggests that humans rule the planet not as much because our brain has more neurons, but *because our language is the most evolved in the animal kingdom*.

This class is a *research* class, which means the grade is focused on your homework, rather than exams. It requires experience in data science and Machine Learning, extensive coding experience with python, and the frame of mind of an explorer. An i5 or higher GPU-equipped laptop is preferrable than CPU-only (computing on Colab or on Cloud TPUs is an option). Each homework will be a new Anaconda notebook where you will take what you learned in class and expand on it, describe your ideas, highlight your computational experiments, and detail your conclusions. In this way, you will learn how to write a research paper. Potentially, you will be able to join a research group with the professor with the goal to write a research paper so that you can talk about your research to highlight your experience in your job interviews.

Attendance/Late work Policy (suggested language, could adjust to course specific content)

Students are expected to complete course readings, participate in class discussions or other learning activities during class, and complete written assignments and research. It is understood that there might be one week when active participation in ongoing class conversations and learning activities might be delayed. Beyond one week time, if there is an absence or lateness in participation (1) faculty must be notified in advance; (2) grades will be adjusted accordingly.

Students must submit assignments by the posted deadline. Students must communicate with faculty prior to the deadline if they anticipate work will be submitted late. Work submitted late without prior communication with faculty will not be considered.

Grading/Evaluation Standards (Required)

Students will be graded based on class participation, homework and research activity.

Course Schedule

Module	**Topic**
1	Topic Models in NLP: Bag of Word (BOW) models and Latent Dirichlet Allocation (LDA)
2	Sequence-to-Sequence transduction architectures with Recurrent Nets (RNNs) and the birth of neural translation
3	Word Embeddings: Converting words to numbers. Learning semantics from syntax
4	The first step: Bahdenau and Luoc Attention with RNNs
5	Game theory with RNNs and a little bit of help from CNNs. Composing Artificial music.
6	Speech to Text models with Connectionist Temporal Classification (CTC)
7	Transformers: Dense networks with Attention everywhere. Better Chinese and Hindi translation with Transformers
8	Beyond NLP: Improving Machine Learning by *refining* NLP models. The few-shot model, BERT, and its descendants
9	Chatbots: Mitsuko, Hugging Face, OpenAI's GPT3, and the one-shot model that terrorizes Elon Musk
10	Emerging technologies: What are Capsule Networks (CapsNets)?

Student Learning/Course Outcomes (SLOs)

By the end of the course, students will be able to build Natural Language Processing (NLP) models and NLP pipelines for automated processing of text with Machine Learning frameworks like Tensorflow and Torch. Students will be able to train a model to translate Chinese and Hindi into English, transcribe voice to text, and improve the accuracy of any Machine Learning model by teaching it language, first. Based on satisfactory completion of this course, a student should be able to conduct research in the field of NLP, apply for Machine Learning coops and jobs specializing in text analysis, and potentially for a more advanced degree based on similar fields of study.

Academic Integrity

A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors.

Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

Student Accommodations

Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability.

For more information, visit http://www.northeastern.edu/drc/getting-started-with-the-drc/.

Library Services

The Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 206,500 e-books, and 70,225 electronic journals.

For more information and for Education specific resources, visit http://subjectguides.lib.neu.edu/edresearch.

Diversity and Inclusion

Northeastern University is committed to equal opportunity, affirmative action, diversity and social justice while building a climate of inclusion on and beyond campus. In the classroom, member of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration and an awareness of global perspectives on social justice.

Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion

TITLE IX

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

In case of an emergency, please call 911.

Please visit www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.