24.S96 Special Seminar:
Linguistically Informed Approaches to Evaluating Neural Network Models*

Course Information

Course Description

Current models in natural language processing are trained on large amounts of text with a simple objective: predict the next word (or predict a word in context). These models have garnered a lot of attention, and there are claims that they can learn non-trivial aspects of human linguistic knowledge. A growing body of literature, framed as “model interpretability”, has attempted to address what exactly such computational models know about linguistic structure. Exploration of these linguistically “naive” models can be used to clarify claims about the nature (and origin) of linguistic knowledge.

In this course, we will survey papers and methods in computational linguistics and natural language processing with an aim towards understanding five key approaches to evaluating neural network models:

- Targeted syntactic evaluations
- Representational probing
- Direct comparison to human behavioral measures
- Priming/fine-tuning
- Cross-linguistic comparison

The course is intended to be a hands-on experience. We will follow a cyclic pattern. First a model will be introduced with a particular approach to evaluation, with students implementing core aspects of both. Then, student led presentations will explore replications, extensions, and challenges to the existing empirical results, broadening our understanding of how to use and evaluate neural models, and how these findings may relate to a theory of language.

Prerequisites

Familiarity with linguistics is expected. No background in programming or machine learning is assumed, but students should come ready to learn relevant skills in a collaborative environment.

---

* This syllabus remains tentative. I reserve the right to modify and/or update this syllabus as the course progresses. You will be notified by email or on canvas of any changes.

Course Materials

All course materials are available for free online or through the library. The course is centered around reading current papers in the field of natural language processing/computational linguistics and its intersection with linguistics. For people new to Python, or programming more generally, I recommend reading and coding along with “Python for Linguists” by Michael Hammond (available through MIT’s libraries here).

Some other useful resources:

- Speech and Language Processing by Dan Jurafsky and James H. Martin
- HuggingFace
- The Missing Semester of Your CS Education
- ACL Anthology

A Note on COVID

Given the continued unpredictability of the COVID pandemic, I understand that things will remain in flux in many of our lives. I am committed to ensuring all students can engage with and learn from this course. Please reach out to me if your situation changes (e.g., you must isolate or quarantine because of COVID) so that we can find a solution that helps you.

Student Learning Outcomes & Objectives

Student Learning Outcomes

- Be familiar with current experimental and evaluation techniques in natural language processing/computational linguistics
- Be able to use some common tools and resources from natural language processing
- Understand ways linguistic knowledge can be learned and represented in large neural models
- Be able to test linguistic hypotheses with state of the art models
- Be able to present papers and experimental results
- Be ready to conduct your own research in model interpretability

Structure

Assignments

There are four main assignments for the course and an initial, smaller assignment (Assignment 0). Each main assignment consists of three parts: 1) a written component, 2) modeling component, 3) evaluation component. The first component is meant to walk you through the core parts of the second and third components by using small examples which can be done by hand. Additionally, you will be asked to reflect on parts of the modeling and evaluation components. The second and third components revolve around implementing aspects of model/evaluation metric in python. You will submit these components via Github classroom. You will have two weeks for each assignment (except Assignment 0). There will be a portion of class time devoted to working on the homework, so that you may ask questions.

Discussion Post

The Sunday before the relevant class meeting, you (or your group) should post some questions, comments, thoughts, etc. you had in reading. These will be collated by the presentator and form the basis of our in class discussion. They should be added to the relevant discussion board on Canvas.

Paper Presentation

In the weeks centered on readings, a student, or a group of students, will be tasked with facilitating the discussion. This should take the form of a short [~20min] presentation of the paper(s). Additionally, the presentator(s) should consider the discussion posts of students and group these questions into some structure which will facilitate discussion. This could take the form of noticing that many questions refer to A, putting questions into dialogue, or whatever natural grouping jumps out to you. You are not expected to prepare answers, rather this is meant to scaffold our class discussion around topics of general interest.

Final Projects

This course culminates in a final project proposal, which can be done alone or in a small group. Note you do not have to have fully implemented your proposed research (though you can if you are feeling inspired). Rather, the goal is to work through how you might evaluate models for your research interests. The proposed project should explore the linguistic knowledge (or lack thereof) of some current neural model of language using techniques discussed in the course. This is broken into 3 components: a topic proposal (~1 page) emailed to me, an in person discussion with me about the project, and a final project proposal (~10 pages). The final project will be graded as follows:

- Topic proposal (5%)
- In person discussion (10%)

- Research question (15%)
- Background literature (30%)
- Explanation of needed resources (eg sample size, resources, model; 10%)
- Explanation of possible findings (30%)

The project should be written up following ACL format and any code should be on github.

Schedule

Date	Topic	Notes	Due Date
9/13	Introductions	In class work: setting up your computer, working with github classroom, basic python exercises and submitting for autograding	Assignment 0 out
9/20	**Intro to Building Models:** Feed-Forward Neural Models for Classification	Lab Day: Introduction to pytorch, small feed-forward neural network for part-of-speech classification, work on Assignment 0	Assignment 0 due Assignment 1 out
9/27	**Word2Vec and Probing**: Assigning vectors to words and learning what they encode	Presentation of Word2Vec and basics of probing Lab: Build Word2Vec and Probing (i.e. work on Assignment 1)	Discussion Post due Sunday 10/2
10/4	**Paper Presentations**: Representational Probing	Ettinger, Elgohary, and Resnik (2016). Probing for semantic evidence of composition by means of simple classification tasks. Hewitt and Liang (2019). Designing and Interpreting Probes with Control Tasks. Voita and Titov (2020). Information-Theoretic Probing with Minimum Description Length. Lovering, Jha, Linzen, and Pavlick (2020). Predicting Inductive Biases of Pre-Trained Models	Assignment 1 due Assignment 2 out
10/11	No Class		
10/18	**RNNs**: Predictions	Presentation of RNNs and LSTMs and basics of targeted syntactic evaluations	Discussion Post due Sunday 10/16

	through time and targeted syntactic evaluations	Lab: Build RNN and TSE (i.e. work on Assignment 2)	Project Proposal Email due
10/25	**Paper Presentations**: Targeted Syntactic Evaluations	Linzen, Dupoux, Goldberg (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. Marvin and Linzen (2018). Targeted Syntactic Evaluation of Language Models. Warstadt, Parrish, Liu, Mohananey, Peng, Want, and Bowman (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs of English. Newman, Ang, Gong, and Hewitt (2021). Refining Targeted Syntactic	Assignment 2 due Assignment 3 out
11/1	**GPT**: Transformer Decoders, Attention, and direct comparisons to human behavior	Presentation of GPT-based models and methods for comparing models to humans Lab: Build GPT and measure surprisal	Discussion Post due Sunday 10/30
11/8	**Paper Presentations**: Comparing models to human behaviors	Wilcox, Gauthier, Hu, Qian, and Levy (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. Schrimpf, Blank, Tuckute, Kauf, Hosseini, Kanwisher, Tenenbaum, and Fedorenko (2021). The neural architecture of language: Integrative modeling converges on predictive processing. van Schijndel and Linzen (2021). Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. Wilcox, Vani, and Levy (2021). A Targeted Assessment of Incremental Processing Neural Language Models and Humans.	Assignment 3 due Assignment 4 out
11/15	**BERT**: Transformer Encoders, Bi-directional Attention, and BPE Tokenizers	Presentation of BERT/RoBERTa based models and BPE Tokenization. Lab: Build RoBERTa and BPE Tokenizer	Discussion Post due Sunday 11/20 Schedule Meeting about Final Project (date established by class)

11/22	**Paper Presentations**: Cross-linguistic comparisons	Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018). Colorless Green Recurrent Networks Dream Hierarchically. Mueller, Nicolai, Petrou-Zeniou, Talmina, Linzen (2020). Cross-Linguistic Syntactic Evaluation of Word Prediction Models. Davis and van Schijndel (2020). Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment. Davis and van Schijndel (2021). Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning.	Assignment 4 due Discussion Post due Sunday 11/27
11/29	**Paper Presentations**: Priming, Fine-tuning, and model abstractions	Prasad, van Schijndel, Linzen (2019). Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. Misra, Ettinger, and Rayz (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. Kodner and Gupta (2020). Overestimation of Syntactic Representations in Neural Language Models. Cho, Chersoni, Hsu, and Huang (2021). Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT.	Discussion Post due Sunday 12/4
12/6	**Paper Presentations**: Ethical and Environmental Issues and Large Language Models	Strubell, Ganesh, and McCallum (2019). Energy and Policy Considerations for Deep Learning in NLP. Bender, Gebru, McMillan-Major, Shmitchel (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 Derczynski, Kirk, Birhana, and Vidgen (2022). Handling and Presenting Harmful Text. Jakesch, Buçinca, Amershi, and Olteanu (2022). How Different Groups Prioritize Ethical Values for Responsible AI.	
12/13	Final Projects	Presentations of Final Project Proposals	

Grading Policy

Percentage	Description
20	Participation and Discussion Posts
5*10	Assignments
30	Final Project
100	Total Possible

Deadlines

The assignments should be submitted by 5pm EST the Friday of the week they are due. For example, Assignment 0 should be completed by Friday September 23. I will subsequently grade the written portion of the assignment over the weekend. The automatic grade score will be used to determine your grade for the coding portion of the assignment. I aim to have your grade posted the following Monday. If you are unhappy with your grade, you can resubmit your assignment. A late penalty of 10% will be accumulated for each week after the initial deadline, but you are free to resubmit as many times as you'd like. Please write to me if you anticipate having trouble submitting your assignment on time. I am more than happy to make accommodations for relevant circumstances!