LING-GA 3340 Seminar in Semantics

Team Project Seminar: Linguistic Knowledge in Reusable Sentence Encoders

Sam Bowman
Tues 3:30–6:15p
10 Washington Place, Rm. 103

This seminar will follow a slightly unusual format: We will write one term paper as a group, with the goals of (i) definitively answering one narrow question in computational linguistics and submitting the results for publication, (ii) giving junior students an introduction to this kind of research, and (iii) giving senior students experience in research mentorship. The seminar will be participant-driven, and the format of the weekly meetings will vary based on what will best allow us to make progress together.

While we will decide on a precise project topic during the first few weeks of the course, I expect the project to fit the broad theme of *linguistic knowledge in reusable sentence encoders*. Work on reusable sentence encoders like ELMo and BERT is built around an ambitious goal: These neural network models are meant to extract representations of sentence structure and sentence meaning that can then be used as the core of the solution to any language understanding problem in NLP. While this goal has not been fully realized, the last year has seen extraordinary progress in this direction, and it increasingly seems that the latest models are understanding sentences in some substantial way. However, as these are neural network-based models with little built-in structure, it is not straightforward to characterize what these models know. The goal of our team project will be to advance our knowledge of this question, with the help of results and concepts from syntax and semantics.

Proposed Theme

> The team project will use the task of acceptability judgment—deciding whether a string of words forms a sentence—to analyze encoders' ability to represent and use syntactic and morphological structure in sentences. We will focus on models that are trained or fine-tuned using the CoLA training set, but we will explore a variety of pretrained encoders (like ELMo and BERT). In service of attaining clear evidence about these models, we will consider a variety of fine-tuning mechanisms (with varying amounts of new parameters at fine-tuning time, and varying amounts of CoLA training data) and a variety of evaluation sets (including both naturally-occurring and custom-written example sentences).

Prerequisites

PhD students *in any program* are welcome to enroll. (Contact me if you need help.) Others should consider taking NLU instead.

You are welcome to participate in the course without formally enrolling if you aren't able to enroll for any reason. If you plan to join the course but can't formally enroll, send me an email before the start of the term so I can add you to the mailing list. *Auditing in the normal sense is not allowed*: Anyone who attends meetings after the end of add-drop (week three) must be committed to contributing to the project and to sticking with the seminar for the full term.

The project will demand several kinds of work, including literature surveys, programming, computational experiment management, statistical analysis, data annotation, and writing. Participants are expected to contribute in several ways, but are not expected to contribute to all aspects of the project. Participants who enter the course without machine learning or programming experience will have to do some extra work to catch up, but are not required to fully acquire these skills during the term to be successful.

Requirements and Grading

We'll agree each week on a breakdown of what each participant is expected to do. Before class each week, each participant should submit a 1–2 sentence update each week on NYU Classes (under Forums) reporting what they've done on the project. This update will be visible to the rest of the group, so contact me privately with any personal concerns.

For students taking the class for credit, I'll use a simple grading scheme meant to make it easy for any substantial contributor to earn an *A*: Weekly updates will be graded on a two-point scale (0 = no contribution, 1 = small contribution, 2 = substantial contribution). 14 points will yield an *A*. A side effect of this is that it's fine to take a few weeks off where you do no work outside of class, though please coordinate this with the rest of the group if you do. I won't grade late submissions without a documented excuse—if you forget a week, just pack two week's worth of updates into your next week's submission (the maximum will still be two points).

Logistics

Day-to-day communication will happen over Slack. Email Sam ASAP if you haven't yet received an invitation.

Office hours are by appointment (default: Monday 3:30–4:30). Contact Sam.

Schedule

1/29: Discussion - Reusable sentence encoders

- Quick review of syllabus
- Read: Peters et al. '18, Devlin et al. '18 - Sam leads discussion
- Assign discussion leaders for next class

2/5: Discussion - Syntax in Sentence Encoders

- Introductions, ~10m

- Read: Warstadt et al. ‘18 (plus presentation of Warstadt and Bowman ‘19) - Alex Warstadt leads, ~40m
- Read: Linzen et al. ‘16, Goldberg ‘19 (plus presentation of Wolf ‘19) - Alicia Parrish leads, ~40m
- Review claimed TODOs, ~10–20m
- Open discussion
- Assign discussion leaders for next class

2/12: Discussion - Analysis Methods

- Report: Tenney et al. ‘19 (plus a survey of analysis methods mentioned there, possibly including Adi et al. ‘16) - Phu Mon Htut leads, ~20m
- Brief intro to Prince - Shikha Bordia leads
- Brief update on initial experiments - Jason Phang and Haokun Liu
- Project discussion: Baseline encoders

2/19: Discussion - Papers relevant to topic of interest; Planning early experiments

- *No required reading – focus on developing your proposed research questions*
- Project discussion/assigning work: Experimental design and infrastructure (1:30m + 40m)
    - Talk through all pending experiment ideas.
    - Goal is not to finalize set of ideas, but rather to make sure we’re laying the groundwork for the kinds of experiment we want to do.
    - For each tentative idea:
        - Do we think it’s important and relevant to the theme of the seminar?
        - Is there a group of people that’s prepared to do it?
        - What infrastructure or data collection work would need to happen before we can start
- Q and A on Prince and jiant (40m)

2/26: Further planning

- Quick updates from everyone who claimed a research question (5m each)
    - Does your question seem viable from what you know now?
    - What needs to happen this week for you to be able to answer your questions this term?
- Presentation of initial baseline results (Haokun/Jason, 15m)
- Brief discussion of GitHub and pull requests (Sam, 10m)
- Brief NAACL paper presentation (Anhad): https://arxiv.org/pdf/1811.00225.pdf (15m)
- Open discussion/revisiting research questions
- Planning for next week

3/5

- General updates.
- Quick updates from everyone who claimed a research question (5m each)

    - What needs to happen this week for you to be able to answer your questions this term?
- Discuss the two NAACL analysis papers (shared privately, available in Drive folder, 15m each; Wei, Shikha)
- Interactive exploration (Shikha, Jason)

3/12

- General updates.
- Quick updates from everyone who claimed a research question (5m each)
    - What needs to happen this week for you to be able to answer your questions this term?
- Possible breakout meetings:
    - Data preparation
    - Code preparation

3/19: Spring Break

3/26: Final planning for experiments

- General updates
- Discuss Wilcox et al. (Hagen, 15m)
- Quick updates from everyone who claimed a research question (5m each)
    - What experiments are you running first? What needs to happen for you to have initial results?
- Coordinate who trains and evaluates which models (15m).

4/2: Pilot results

- General updates
- Substantial updates from everyone who claimed a research question (~15m each)
    - What results do you have so far? What follow-up experiments do you want to do in light of these results?
- What new code do we need to make everything easily reproducible? (15m)

4/9: Initial full set of results

- General updates
- Substantial updates from everyone who claimed a research question (~10m each)
    - What are your initial conclusions? What further analysis do you need to do?
- General discussion:
    - Which research questions should our group paper deal with?
    - Which research questions could form good separate papers?
    - Are there any research questions we should drop?

4/16: Paper planning

- Framing the main paper (Yining, Alex, Jason, ~30m)

- Outlining all three papers, assigning sections (~1h)
- Brief overview of code status (Sam, ~15m)
- Discussion of specific experiments (~30m)
    - What's left to do to finalize our results?
    - What's left to do to make our results easily reproducible?
    - What follow-up analysis should we do?

4/23: Debugging, analysis, and follow-up experiments

- Discuss Hewitt and Manning '19 (Shikha, ~10m)
- Updates, potentially including initial results, and including discussion of presentation mechanisms (~15m each)
    - Framing/literature review
    - Data creation
    - Base pretrained models
    - Main NPI/CoLA experiments
    - Probing
    - Cloze/MLM
- Structural probing (not NPIs)
- Analysis and further questions (50m)

4/30: Writing and visualization

- *Deadline: Draft version of all results and analysis in paper*
- Updates, **full results**, and including discussion of presentation mechanisms (~10m each)
    - Introduction/abstract
    - Literature review
    - Data creation
    - Base pretrained models
    - Main NPI/CoLA experiments
    - Probing
    - Cloze/MLM
    - Discussion/conclusion
    - *Transformer Attention Paper (quick overview of full paper plan, 15m)*
- Reruns and new visuals
- Timeline for finishing and planning for next week (15m)

5/7: Writing and visualization

- *Deadline: Complete readable draft*
- *ICLR conference - SB Away*
    - *Timing czar (chosen randomly): Yining*
- Group editing by section: Talk through the draft paper (~15m each, transformer attention offline)
    - Introduction/abstract

- Literature review
- Data creation
- Base pretrained models
- Main NPI/CoLA experiments
- Probing
- Cloze/MLM
- Discussion/conclusion

(EMNLP paper deadline is May 21—with abstracts due the week before—next TACL deadline is June 1)

Resources

We'll be building on the jiant codebase, which is built on AllenNLP, PyTorch, and Python 3.

Sick Days

If you're sick, please don't come to class. I won't count your participation. If it's your day to present something in class, try to warn me, and try find a way to trade with someone (and send them any notes you have that are clear enough to share). Even if that doesn't work out, though, stay home.

Applicable University Policies

Academic Integrity

Work you submit should be your own. Please consult the CAS academic integrity policy for more information: http://cas.nyu.edu/page/academicintegrity. Penalties for violations of academic integrity may include failure of the course, suspension from the University, or even expulsion.

Religious Observance

As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. The policy and principles to be followed by students and faculty may be found here: The University Calendar Policy on Religious Holidays (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html)

Disability Disclosure Statement

Academic accommodations are available to any student with a chronic, psychological, visual, mobility, learning disability, or who is deaf or hard of hearing. Students should please register with the Moses Center for Students with Disabilities at 212-998-4980.

NYU's Henry and Lucy Moses Center for Students with Disabilities

726 Broadway, 2nd Floor
New York, NY 10003-6675
Telephone: 212-998-4980
Voice/TTY Fax: 212-995-4114
Web site: http://www.nyu.edu/csd