Notes

CS 685, Fall 2025
Advanced Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

- Extra credit talk writeups
  - New requirement: submit writeup **within 3 days of the talk**
  - Two opportunities this week

---------- Forwarded message ---------
From: **Tessa Masis** <tmasis@umass.edu>
Date: Thu, Nov 13, 2025 at 9:19 AM
Subject: NLP Seminar 11/19/25: Shira Wein
To: <umass-nlp@googlegroups.com>, <cics.mlfl@gmail.com>, <seminars@cs.umass.edu>, <ling-dept@linguist.umass.edu>

Hello everyone! We have our last NLP seminar talk next week, and we're excited to have Shira Wein speaking on **Lost in Translation, and Found: Detecting and Interpreting Translation Effect in LLMs** from 12:30-1:45pm on Wednesday, November 19 in LGRC A104A. We hope to see you there!

Shira will be coming to speak in person, so there will be a few one-on-one meeting slots available with her - please message me if you're interested!

Presenter: Shira Wein (in-person talk)

When: Wed, Nov. 19, 12:30-1:45pm

Location: LGRC A104A

Zoom link: https://umass-amherst.zoom.us/j/93786229674?pwd=8xoWcmL10ZgznXWugmH7CwyYaDp26w.1

Title: Lost in Translation, and Found: Detecting and Interpreting Translation Effect in LLMs

Abstract: Translated texts bear several hallmarks distinct from texts originating in the language (“translationese”). Though individual translated texts are often fluent and preserve meaning, at a large scale, the presence of translated texts in training data negatively impacts performance and in test data inflates evaluation. In this line of work, I investigate (1) whether humans are able to distinguish texts originally written in English from texts translated into English, (2) how the surface-level features of translationese can be mitigated using Abstract Meaning Representation, and (3) why neural classifiers are able to distinguish original and translated English texts much more accurately than humans.

Bio: Shira Wein is an Assistant Professor of Computer Science at Amherst College, where she works on multilingual natural language processing. Shira is specifically interested in computational semantics, multilingual data, and model evaluation practices. Before joining Amherst in 2024, Shira received her Ph.D. in Computer Science from Georgetown University, and completed research at Google, the University of Southern California, and the NASA Jet Propulsion Laboratory. In addition to research, Shira is passionate about teaching courses across the computer science curriculum and mentoring undergraduates.

Related Papers:

- Lost in Translationese? Reducing translation effect using Abstract Meaning Representation (Wein & Schneider, EACL 2024)
- Human raters cannot distinguish English translations from original English texts (Wein, EMNLP 2023)

---------- Forwarded message ---------
From: **Rangel Daroya** <rdaroya@umass.edu>
Date: Mon, Nov 17, 2025 at 8:38 AM
Subject: [MLFL] Thu 11/20 12-1pm, Yuanqi Du (Cornell)
To: <seminars@cs.umass.edu>
Cc: <cics.mlfl@gmail.com>

Hi everyone,

This week at Machine Learning and Friends Lunch we will host Yuanqi Du, a PhD candidate in Computer Science at Cornell University. His research is on developing principled, efficient probabilistic and geometric models that accelerate scientific discovery, from hypothesis search, validation to automation, with a special focus on the intersection of physics and chemistry and their applications in drug and materials discovery.

If you are interested in a research meeting with Yuanqi Du, please sign up in this Google doc. Please feel free to email cics.mlfl@gmail.com for any questions and comments.

who: Yuanqi Du (https://yuanqidu.github.io/)
when: Thursday, 11/20/2025, 12pm-1pm
where: CS 150/151 (pizzas available), Zoom
food: Pizza and drinks

Title: Scientific Knowledge Emerges in LLMs and You Can Extract It

Abstract:
The emerging capabilities of large language models (LLMs) are opening new frontiers in scientific research, including experiment operation, literature retrieval, and molecular design. A central question, however, is whether LLMs truly encode scientific knowledge—and if so, how this knowledge can be systematically extracted. In this talk, I will present an affirmative answer to this question, supported by strong quantitative and empirical evidence. I will begin by framing knowledge extraction as a search problem with a computational verifier. I will illustrate through three problems: molecular optimization, crystal structure generation, and retrosynthesis. In all three cases, LLMs demonstrate impressive performance compared to state-of-the-art computational approaches. I will conclude by reflecting on analogous discoveries in other scientific domains and highlighting key questions for future exploration.

Bio:
Yuanqi Du is a PhD candidate in Computer Science at Cornell University, where he studies the intersection of AI and scientific discovery. His research centers on developing principled, efficient probabilistic and geometric models that accelerate scientific discovery, from hypothesis search, validation to automation, with a special focus on the intersection of physics and chemistry and their applications in drug and materials discovery. Yuanqi's work has appeared in leading machine learning venues (NeurIPS, ICML, ICLR) and featured as cover articles in top-tier scientific journals, including Nature, Nature Machine Intelligence, Nature Computational Science, and Journal of the American Chemical Society. As a passionate community builder, Yuanqi has organized over 20 community events, including conferences, workshops, and seminars across AI for Science, geometric deep learning and probabilistic machine learning.

---------- Forwarded message ---------
From: **Krista Gile** <gile@umass.edu>
Date: Mon, Nov 17, 2025 at 8:52 AM
Subject: [CSSI-events] This Friday, November 17 | Matthias Roesti, "From RAGs to riches? An efficient pipeline for exploratory text analysis"
To: <CSSI-events@cs.umass.edu>

Please join us this Friday for a Computational Social Science Institute (CSSI) lunch and research seminar, jointly hosted with DACSS.

Friday October 21, 12pm-1:30pm
Lederle Graduate Research Center, Room A112

[RSVP Form (encouraged, not required, also allows you to request meeting with speaker)]

Speaker: Matthias Roesti, Research Associate at the University of St. Gallen, Switzerland, visiting Brown University

From RAGs to riches? An efficient pipeline for exploratory text analysis

Abstract: I introduce a scalable computational approach for comparing large bodies of text across many research questions using retrieval-augmented generation (RAG). By combining language models with domain-specific retrieval, the method enables fast, flexible, and customizable text analysis - offering a powerful alternative to traditional document-by-document approaches.

Bio: Matthias Roesti is a Visiting Scholar at Brown University’s Department of Economics and a Postdoctoral Research Associate at the University of St. Gallen. His research combines political economy and computational social science to examine how digital information systems and organized interests shape public discourse, using large-scale text, media, and algorithmic data.

Lunch will be provided.
Lederle Graduate Research Center (LGRC) is located at 740 N. Pleasant Street.

{}

As always, please share this information with others who might be interested, and encourage them to email us if they want to get involved and become CSSI affiliates. Upcoming events are listed at http://www.cssi.umass.edu/events and anyone can sign up for the CSSI-Events mailing list.