Advanced Topics in Computational Semantics

Overview of Research Projects

Martha Lewis

ILLC
University of Amsterdam

4 April 2025

Outline.

Outline.

Recap

Last time, we discussed:

- Overall course organization
- Early models of computational semantics
- Representation Learning and Deep Learning
- Overview of the practical

Today, we will look at the research projects in more detail.

What we will do today

- Opportunity for questions: are you clear where to look for information about the course?
- Research projects in a little more detail
- How to carry out a good research project

Outline.

Research project topics

1. Large Language Models and Group Fairness
2. Stereotypes in Language (Models)
3. Universal Emotion Embeddings
4. Misinformation and Disinformation

Submit your top three choices on Canvas by **Monday, 14th April**

Topic 1: Large Language Models and Group Fairness

LLMs have achieved much success in NLP, including translation, writing code, answering healthcare questions, and can be used to detect bias and abusive language. However:

- Not everyone benefits equally
- Benefits are not equitably distributed across groups
- This happens even when group is not explicitly stated
- I.e., models will detect groups via language used

We want models to treat people from different groups equitably
There are two projects on this topic.

Project 1

Obtaining empirical evidence for group (un)fairness

- Collect data from humans on prompts for task solving with LLMs
- Analyze the prompts to detect whether they differ across demographics
- Obtain responses and internal representations from LLMs
- Train a classifier on the internal representations to discover whether the LLM encodes prompts from different demographics differently.

Project 2

Project 2: Using LLMs to improve bias detection

- Use an LLM to detect bias in Dutch government documents
- This involves prompting and fine-tuning open-source LLMs using parameter efficient methods.
- Compare accuracy of bias detection to see how bias detection differs across groups.
- Use feature attribution methods to attribute models predictions to specific terms.

Topic 2: Stereotypes in Language (Models)

Both humans and LLMs perpetuate stereotypes through language: compare 'he is assertive' vs 'she is aggressive'

- The kinds of bias that humans display have been well studied.
- Differences in the concreteness and abstractness of terms people use (Linguistic Intergroup Bias)
- Differences in sentiment when describing in-group vs out-group behaviours.
- Differences in context and explanation given (Stereotypic Explanatory Bias).

We want to detect and understand biased behaviour from LLMs

Types of Bias

Linguistic intergroup bias (LIB)

- Humans use abstract terms for stereotypical behaviour: ‘he is smart’
- ... and concrete terms for counter-stereotypical behaviour: ‘she did well on the test’

Stereotypic Explanatory Bias (SEB)

- Humans use short descriptions for stereotypical behaviour: ‘she is crying’
- ... and add explanations for counter-stereotypical behaviour: ‘he is crying because he had a bad day’

Project 1

Linguistic intergroup bias (LIB)

- Set of stereotypical and counter-stereotypical descriptions of behaviour.
- Use these as prompts for an LLM
- Analyze whether output follows the same patterns as humans.
- Swap the assignment of behaviours between in-group and out-group and analyze changes in descriptions of behaviour.

Project 2

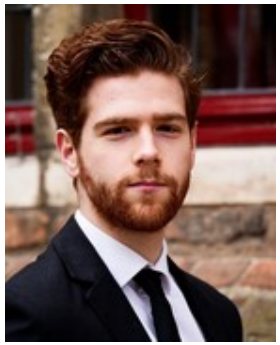
Stereotypic Explanatory Bias (SEB)

- Same dataset of behaviours
- Examine outputs of LLMs explicitly trained for reasoning.
- Do LLMs give more explanation for counterstereotypical behaviour?
- You can also look at concrete/abstract terms in this project, and vice-versa

Both projects also have datasets in multiple languages, so you can compare across languages as well!

Topic 3: Universal Emotion Embeddings

Emotion recognition has been a goal of artificial intelligence research since the very beginning of the field.



- Emotion classification is very difficult.
- There are many different theoretical frameworks for emotion.
- Datasets have different labels and little overlap.
- Can't test generalization effectively.

How can we build and test models that generalize?

Project 1

Supervised Contrastive Learning for Universal Emotion Embeddings

- Standard representation learning methods require similarity of labels
- Contrastive representation learning requires only a means of sampling similar and dissimilar examples
- It can be used to learn with examples from different datasets
- Representations transfer to unseen datasets.

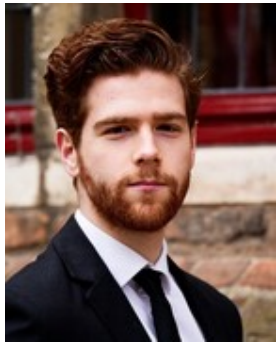
Project 2

Predicting OoD Generalisation for Intermediate Dataset Transfer

- Fine-tuning on an emotion dataset.
- Transfer to another downstream emotion dataset.
- Discover structure within different datasets.
- Predict transfer performance between emotion labelling schemes.

Topic 4: Misinformation and Disinformation

Online misinformation has gone from being a nuisance to being part of our new political reality.



- Social media allow easy sharing of fake news.
- Automated content moderators are not reliable.
- Expert fact checkers cannot keep up with the volume of content produced.
- Misinformation vs. disinformation

How can we detect and combat dis/misinformation?

Project 1

Effect of Propaganda on LLMs

- LLMs are good generators of disinformation
- Ability to generate large amoutns of text with minimal effort is a serious societal threat
- Test the effects of propaganda techniques on LLM generations.
- Use a causal inference framework to understand the effect on LLM output.

Project 2

Meta-learning for Emerging Misinformation Detection

- Emerging news is particularly difficult to classify as fake or real.
- Large-scale dataset to simulate emerging misinformation scenarios.
- Use episodic learning to train a classifier.
- Training various meta-learners in an effort to build adaptable misinformation detectors.

Outline.

Working on a research project

Key steps

1. Formulate your goal or research question
2. Choose methods / models to use
3. Design experiments to test the methods (datasets, baselines)
4. Conduct evaluation: compare the models in terms of performance (quantitative results)
5. Conduct qualitative analysis

Getting started

Project topics come with brief project descriptions on Canvas and some suggested literature

1. read the papers on the topic
2. look at the available datasets
3. find out what the state-of-the-art model is for your task
4. build on top of this state-of-the-art model
   - sometimes there can be several types of models (near-SOTA)
   - numbers alone should be taken with a grain of salt
5. use ideas and models studied in the course, and research wider literature

Designing experiments

1. Choose your baselines wisely:
   - make sure the models are **comparable**
   - a good baseline model does everything the way your model does, except for **the one thing** that you are evaluating
2. Perform ablation experiments:
   - add one technique at a time
   - determine its contribution
3. Compare to prior research (when possible)

Training and evaluation: good research practice

- **Training, development** and **test** splits
  - development set used for parameter tuning
  - test set kept unseen!
  - use standard split, if available in the literature
- **Cross-validation**
  - a viable alternative for smaller datasets
  - use stratification
  - standard dataset splits may be available
- If performance differences are small: use **statistical significance testing**

Conducting experiments: the reality

You came up with your brilliant idea!

You have performed all of the above steps perfectly!

And yet... it doesn't work...

What do you do next?

Conducting experiments: the reality

You came up with your brilliant idea!

You have performed all of the above steps perfectly!

And yet... it doesn't work...

What do you do next?

Not this...

Also not this...

You do this

- Try to diagnose the problem
  - look at the data, perform **error analysis**
  - play with **parameter settings**
  - conduct an experiment under "ideal conditions":
    e.g. equal dataset sizes in a multitask learning setup
  - also **talk to us** at this point!
- Change your setup and try again
  - experiment with a **different dataset**
  - experiment with **variants of the model**, or a different architecture
- Getting a positive result often requires several iterations!

You do this

- Try to diagnose the problem
  - look at the data, perform **error analysis**
  - play with **parameter settings**
  - conduct an experiment under "ideal conditions":
    e.g. equal dataset sizes in a multitask learning setup
  - also **talk to us** at this point!
- Change your setup and try again
  - experiment with a **different dataset**
  - experiment with **variants of the model**, or a different architecture
- Getting a positive result often requires several iterations!

You do this

- Try to diagnose the problem
  - look at the data, perform **error analysis**
  - play with **parameter settings**
  - conduct an experiment under "ideal conditions":
    e.g. equal dataset sizes in a multitask learning setup
  - also **talk to us** at this point!
- Change your setup and try again
  - experiment with a **different dataset**
  - experiment with **variants of the model**, or a different architecture
- Getting a positive result often requires several iterations!

Conducting an analysis

1. Find ways to visualise different aspects of your model
   - e.g. graphs, tSNE plots etc
2. Investigate model behaviour under different conditions
   - e.g. the effect of training data size
   - or performance across different classes
3. Qualitative analysis
   - perform error analysis
   - what does your model do well and where does it fail
   - other interesting trends that the data shows

Conducting an analysis

1. Find ways to visualise different aspects of your model
   - e.g. graphs, tSNE plots etc
2. Investigate model behaviour under different conditions
   - e.g. the effect of training data size
   - or performance across different classes
3. Qualitative analysis
   - perform error analysis
   - what does your model do well and where does it fail
   - other interesting trends that the data shows

Conducting an analysis

1. Find ways to visualise different aspects of your model
   - e.g. graphs, tSNE plots etc
2. Investigate model behaviour under different conditions
   - e.g. the effect of training data size
   - or performance across different classes
3. Qualitative analysis
   - perform error analysis
   - what does your model do well and where does it fail
   - other interesting trends that the data shows

Outline.

Summary

What we covered today

1. We went over the content of the projects: 4 topics, 2 projects per topic
   - Large Language Models and Group Fairness
   - Stereotypes in Language (Models)
   - Universal Emotion Embeddings
   - Misinformation and Disinformation

Summary

What we covered today

1. How to approach a research topic
   - Basics: getting started, designing experiments, training and evaluation.
   - Sometimes (often!) the experiment does not go as expected.
   - What should you do? Diagnose the problem, change the setup, look at your results in different ways: visualization, error analysis.

Coming next...

On Tuesday:

- Lecture: Attention and Transformers (Ivo Verhoeven)

On Friday:

- Seminar: The BERT model