096222: Language, Computation and Cognition
Project Guidelines

2 May 2023

Students enrolled in 096222 are required to complete a class project which constitutes 50% of the course grade. The project is due **July 2nd** at 23:55. You will work on the project in pairs. Groups of three are allowed, with the expectation of an **extended project scope**.

Project Topics You will carry out one of the two projects proposed below, which are based on assignments from homeworks 2 and 3. Both project options consist of three parts - a structured task, a semi-structured task, and an open-ended task. The open ended task is the primary component of the project.

Exceptions In a small number of exceptional cases, students can propose a different project instead. Such a project should build on the learning and assignments work you've been doing in this subject during the semester, and include a substantial computational modeling component (understood broadly). We also encourage coupling your computational modeling work with empirical analysis of some linguistic dataset and/or with behavioral experiments (e.g., through Mechanical Turk).

All projects should include a written report, which should follow from the content of the project. The length should be around 5-7 pages, roughly comparable to a 6 pages proceedings paper for Cognitive Science – the Cognitive Science Society conference / ACL – Association for Computational Linguistics. Please use the Cognitive Science[1] or ACL[2] conference templates for your writeup. Make sure you provide enough plain-English context that it is easy for us to understand your work, and that the key scientific and/or engineering questions are clear. Figures are generally helpful to include in your writeup. You are also welcome to provide a link to a GitHub repository with your code, but you cannot expect the reader to consult your raw code in understanding your writeup.

[1] https://cognitivesciencesociety.org/submissions/

[2] https://2021.aclweb.org/calls/papers/#paper-submission-and-templates

Timeline

- **by May 14 (23:55)** team up with a partner and choose a project. For project options 1 and 2: register on Moodle your team members and chosen project. Students who wish to work on their own project: register on Moodle your team members, and send us by email a 1-page project proposal, and your availability for a (20 minutes) meeting on Tuesday May 16.
- **by June 4 (23:55)** all projects: submit on Moodle a 1 page progress report. It should specify your progress on the structured and semi-stuctured tasks, and a plan for the open-ended task. Projects in good standing should have the structured and semi-structred tasks completed or nearly completed. Additionally, you will be asked to specify your availability for a 15 minutes progress update meeting on Tuesday June 6 / Thursday June 8.
- **by July 2 (23:55)** Submit the final report.

Late Submissions Policy

Projects submitted after the deadline will be deducted 3 points per day, up to a maximum of 7 days. Projects submitted over a week after the deadline will not be graded.

Project 1: Surprisal and RTs

In Homework Assignment 2 you examined the main result from Smith and Levy (2013) regarding the relation between reading times and surprisal using n-gram surprisals that we provided. In this project you will extend this work in a number of ways.

Structured Tasks (15 points)

Train an RNN (LSTM) language model

We ask you to train the RNN on a section of the Penn Treebank, a commonly used dataset in natural language processing research. We have provided the files `ptb_tok_train.txt` and `ptb_tok_dev.txt`, which are the training and validation sets, respectively.

Please see the Colab notebook[3] for detailed instructions on how to train your RNN language model. Note: training may take up to 2 hours, though Colab compute speed is unpredictable and training may finish in as little as 20 minutes.[4]

Obtaining surprisals. Once you have trained the RNN, you need to obtain surprisals from the model. We have also provided this code for you. Please see the Colab notebook for detailed instructions.

Alignment. Next, you'll need both the RNN model surprisals and n-gram model surprisals. We leave the alignment of the surprisals derived from these two models, to the reading times, for you to complete. The `harmonize.py` function you wrote in Homework 2 will be helpful here.

Compare n-gram and RNN models

Next, let's zoom in on the difference between the surprisals derived from the n-gram model and the RNN model. In the following analyses it is sufficient to use a univariate model without control predictors.

1. Which model has surprisal estimates which correlate better with human reading times?

2. Plot the relationship between the n-gram model's surprisal estimate for a word and the RNN's estimate. Each point in the graph should correspond to a single token from the harmonized data. Describe what you see in this graph. Are the models generally well matched? On what parts of the surprisal spectrum do they disagree?

3. Pick specific interesting points from this graph (for example, points where the two models have very different surprisal estimates) and report the sentences containing the corresponding tokens. Why do you think the models should disagree?

[3] https://colab.research.google.com/drive/1ybProvI3-eSE42dsPM1zIHzGWWcNMWbD?usp=sharing

[4] 2 hours is very fast by NLP standards, and it will be running on the cloud, not your local computer. :)

4. Examine **spillover** in both models: look at the relationship between word probability and the *next* word's reading time (a "spillover" effect). Is the effect similar as on the current word? Bigger? Smaller? Different shape? Are the spillover effects different across the two models?

Semi-structured Tasks (25 points)

Choose (at least) two of the following tasks. Groups of three students must do all three tasks.

- Fit and plot the RT surprisal curve using a General Additive Model (GAM). The model should include control variables for log-frequency and word length. Examine both current word and spillover effects.
- Choose a different reading times corpus (self-paced reading, eye-tracking, or maze), compute surprisals for this corpus using the RNN model from the structured task, and analyze the RT-surprisal relationship, including spillover. Compare the results to those you obtained with the self-paced reading corpus of Smith and Levy (2013) used in HW2 and the structured task.
- Train n-gram and RNN models on a larger dataset, such as Wikitext-2, Wikitext-103, and/or a still larger dataset (note that training larger models on Colab might require some careful checkpointing work), and see what happens to the relationship with RTs.

The code we provided for the structured task may be helpful here too. We recommend going over the code!

Open-ended Task (60 points)

Devise and carry out an additional substantial analysis not listed above.

Resources

The following tools and datasets may be useful for this project, in particular for the open-ended task. Note that you are encouraged, but *not required* to use them, and are further encouraged to explore other relevant resources and papers.

Tools:

- Minicons, LM-ZOO: Packages for extracting contextualized word vectors and surprisals from various language models.
- KenLM: Package for training n-gram models.
- HuggingFace: Easy access to a large variety of NLP models.

Datasets:

- **EEG** UCL Corpus (Frank et al., 2015). Please be sure to read Frank et al., 2015 if you use this data, and note that this corpus also has Eyetracking and SPR data (ESM1) (Frank et al., 2013).
- **Eyetracking** GECO (Cop et al., 2017), MECO L1 (Siegelman et al., 2022), MECO L2 (Kuperman et al., 2023), CELER (Download) (Berzak et al., 2022), PROVO (Eyetracking + Cloze) (Luke & Christianson, 2018). Additional eyetracking datasets, including in languages other than English can be found here.
- **Self Paced Reading** Natural Stories SPR (Futrell et al., 2021)
- **Maze** Natural Stories Maze (Boyce & Levy, 2023)

Project 2 is on the next page.

Project 2: Word Embeddings and the Brain

In Homework Assignment 3 you replicated Analysis 1 from Pereira et al., 2018 on decoding words from fMRI data. In this project you will extend this work as follows.

Structured Task (15 Points)

Sentence decoding

- Perform the analysis of Homework Assignment 3 question 3 using another type of static word embeddings (e.g. Word2vec) and compare the results to those you obtained with GloVe.
- Read Pereira et al., 2018 and describe the similarities and differences between analyzes 1, 2, and 3 in that paper.
- Use the GloVe based decoder model you trained in Homework Assignment 3 question 3 and test it on the datasets from analyses 2 & 3. Each dataset contains sentence representations (i.e. a vector representation averaged over all the words in the sentence) and the corresponding neural data from an individual subject (384 sentences from analysis 2 and 243 from analysis 3; The datasets are available in a Google Drive Folder[5]). For each dataset, use the learned decoder model to decode sentence representations and evaluate the performance via the rank accuracy method (as you did in HW3).
- Each sentence, in both datasets, is related to a specific passage (a single passage contains 3 or 4 sentences), and every passage is related to a specific broad topic (e.g., musical instrument, animals, etc. The labels for the sentences/passages are available in the Google Drive folder as well). You will need to analyze the accuracy scores from the previous section and try to identify the topics where the decoder was more / less successful in predicting the sentences.

Semi-structured Tasks (25 Points)

Perform the following two tasks

- Train a decoder model on either the dataset from analysis 2 (384 sentences) or from analysis 3 (243 sentences) using both (1) the sentence representations that were used in the paper (the same representations from the structured task) and (2) sentence representations as extracted from a contextualized word embedding model (such as BERT, GPT2, GPT3, etc.). Report and compare the results from both methods.
- Build a **brain-encoder** model. Instead of predicting sentence identities using neural signals (i.e., neural **decoding**), you will try to predict human neural signals from the embedding vectors representations of the sentences (neural **encoding**; you can read

[5]https://drive.google.com/drive/folders/1cwciPYnnmPEReE0tpX78SQlqlwL88V8b?usp=sharing

about neural encoder in Huth et al., 2016's paper). We ask you to fit a separate linear-regression model for each voxel in the dataset related to analysis 2 (384 sentences) or 3 (243 sentences) of Pereira et al., 2018 (180 concepts). For each voxel/model, calculate the R^2 score and examine how many voxels are *significantly* associated with the information embedded in the word vectors, and how well those voxels are predicted. This analysis should be run twice: once using the non contextualized vector representations (The original vector representations from the paper), and another time, using the contextualized representations you extracted before.

Open-ended Task (60 Points)

Carry out an additional substantial analysis not listed above.

Resources

The following tools and datasets may be useful for this project, in particular for the open-ended task. Note that you are encouraged, but *not required* to use them, and are further encouraged to explore other relevant resources and papers.

Tools:

- HuggingFace: Easy access to a large variety of NLP models.

fMRI Datasets:

- Natural Stories (Shain et al., 2020).
- Tang et al., 2023 Data and Code.
- Mitchell et al., 2008 Nouns Dataset. Please contact Refael if you would like to use this dataset.

References

Berzak, Y., Nakamura, C., Smith, A., Weng, E., Katz, B., Flynn, S., & Levy, R. (2022). Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, *6*, 41–50.

Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*, *2*(1).

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, *49*, 602–615.

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, *45*, 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The erp response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2023). Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, *45*(1), 3–37.

Luke, S. G., & Christianson, K. (2018). The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, *50*, 826–833.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191–1195.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1–13.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). Fmri reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2022). Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 1–21.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 2022–09.