Finding datasets / resources

LING575 Analyzing Neural Language Models
Shane Steinert-Threlkeld
April 20 2022

Roles for data

- You will need data for your analysis project
- One simple case: data captures linguistic feature X, ask which representations in which models can capture that feature
  - (Can be good to use more than one dataset here if possible)
- More complicated: generate your own data
  - Because you hypothesize that model X will struggle with it (“adversarial”)
  - To carefully control various linguistic variables
    - Can borrow / take inspiration from / build upon examples from linguistics papers
  - Examples: Marvin and Linzen 2018, Warstadt et al 2019, McCoy et al 2019

What makes a good dataset?

- Can depend on the project; try to find/build data that's motivated by your question/hypothesis
- Well-designed:
  - Clear annotation guidelines that yield consistent results
  - Targets the intended task
- Relatively large (somewhat less important for analysis projects)
- Precedent in the literature
  - If your project involves phenomena that are well-studied in NLP, use (and/or compare with) existing datasets!
  - Can, e.g., be a new analysis using data from a paper we've already discussed

LDC; Treehouse DB

- The Linguistics Data Consortium has many excellent datasets (think Penn Treebank)
- Many of those, and lots more, pre-installed on paths
  - For a complete directory, see https://cldb.ling.washington.edu/

SemEval

- International Workshop on Semantic Evaluation
- Each year, a shared task (or tasks)
  - Multiple teams build models for one task
  - Data is well-designed to be consumable by teams
- 2022 (links to older): https://semeval.github.io/SemEval2022/
- Not every task will be appropriate; but you can search for your keywords + “semeval” and see if there’s been a task in the past
- NB: there are other shared tasks, not just SemEval, so you can also try keywords + “shared task”

Some general resources

- HuggingFace datasets hub:
  - https://huggingface.co/datasets
  - Good coverage of commonly used benchmarks; nice inter-operability with transformers library
  - Less coverage of adversarial / smaller / targeted datasets
- New-ish: Google Dataset Search
  - https://datasetsearch.research.google.com/
  - Personally some mixed results so far, but could be very useful
- The Big Bad NLP Database
  - https://quantumstat.com/dataset/dataset.html
  - New, has large/standard datasets, but fairly small coverage (low recall)

Some Particularly “Linguistic” Datasets

- CoLA (acceptability): https://nyu-mll.github.io/CoLA/
- BLiMP (minimal pairs, many phenomena; artificially generated with decent vocab): https://doi.org/10.1162/tacl_a_00321 , https://github.com/alexwarstadt/blimp
- NOPE (natural presuppositions): http://dx.doi.org/10.18653/v1/2021.conll-1.28 , https://github.com/nyu-mll/nope
- Decompositional Semantics: decomp.io
  - Rachel Rudinger’s guest lecture last year is available on Canvas!
  - Large-scale annotations (simple framework) for many phenomena: factuality, time, semantic proto-roles, …
- EntailmentBank (open domain): https://allenai.org/data/entailmentbank

Special Topics Presentations

Presentations

- Each group will be responsible for leading an ~45 minute discussion on a special topic of their choosing
- For example:
  - A deep dive into one or two papers that are important to your group's project
  - Survey of a method / model / dataset that you are using that was not covered in the earlier lectures
- Present material, but also lead/guide a discussion, to make these sessions as much seminar-style as possible
  - You don't need to have all the answers about everything that could possibly come up

Logistics

- Sign up here:
  - https://docs.google.com/spreadsheets/d/1Z_Qjk4A_T_EBwG0_SjMZmUQw5yg6JQoKLmzByvJkrJs/edit?usp=sharing
  - For now: pick a time slot. You only need to fill in the first two columns.
  - NB: there are 7 groups; so one week will have only one presentation
- **One full week before your presentation:**
  - Fill in topic, and list of reading(s) / resources
  - Email me as well
  - I will post to the website so that everyone can read in advance

Next time

- Some tips / advice about
  - Managing projects
  - Writing papers
- Useful resources / libraries
- Walkthrough of basic diagnostic classifier example