CS5740: Natural Language Processing

Meta NLP

Instructor: Yoav Artzi

Slides inspired and adapted from Sam Bowman

Overview

- Starting your NLP project
  – Literature, data, evaluation, and hyperparameter tuning
- Evaluating your work
- A Few Words on Ethics in NLP

Literature Review

- Do this early!
- Why?
  - Don’t re-invent the wheel!
  - Learn about common tricks, resources, and libraries that will make your life easier

Literature Review

- Identifying papers:
  1. Do a keyword search on Google Scholar, Semantic Scholar, or the ACL Anthology
  2. Download the papers that seem most relevant
  3. Skim the abstracts, intros, and previous work sections
  4. Identify papers that look relevant, appear often, or have lots of citations on Google Scholar
  5. Download those papers
  6. Return to step 3

h/t Bill MacCartney

Literature Review

- Where to find the most trustworthy papers:
  - NLP: Proceedings of ACL conferences (ACL, NAACL, EACL, EMNLP, CoNLL, LREC), Journal of Computational Linguistics, TACL, COLING, arXiv*
  - Machine Learning/AI: Proceedings of NIPS, ICML, ICLR, AAAI, IJCAI, and arXiv*
  - Computational Linguistics: Journals like Linguistic Inquiry, NLLT, Semantics and Pragmatics
- What to mention:
  - General problem/task definition
  - Relevant methods and results
  - Comparisons with your work and other related work
  - Open issues

Acquiring Datasets

- Find existing datasets
  - ACL anthology
  - Linguistic Data Consortium (LDC)
  - Look for datasheets when available (e.g., QuAC)
- Find them in the wild
  - Example: Ubuntu Dialogue Corpus, StackOverlow Data
  - Careful: easy to violate copyright and terms of service!
- Build them
  - Write detailed guidelines, and work with experts
  - Write simple guidelines, and crowdsource
- Generate them
  - Super easy
  - Controlled → help analysis
  - Careful: artificial data does not reflect the real world!

Quantitative Evaluation

- Follow prior work and use existing metrics
  - New task? Create a metric before you start testing. It must be independent of your model!
- Use ablations to study the effectiveness of your choices (and don’t adopt fancy solutions that don’t really help)
  - Example:
    - MLP sentiment classifier with GloVe embeddings
    - MLP sentiment classifier with random embeddings
    - MaxEnt classifier with GloVe embeddings
    - MaxEnt classifier with random embeddings
- Consider controlled human evaluation when standard, and even when less standard
  - Example: summarization, machine translation, generation
- Test for statistical significance when differences are small and models are complex
- Consider extrinsic evaluation on downstream tasks
- Negative results are informative too!

Qualitative Evaluation and Error Analysis

- Your goal: convince that your hypothesis is correct
- Interesting hypotheses are often hard to evaluate with standard/intuitive quantitative metrics
  - Example: Attention-based NMT models can learn the same kinds of alignments as phrase-based MT systems, but generalize better to unfamiliar words and phrases.
- Qualitative evidence can help!
- How to start?
  - Look to prior work!
  - Show examples of system output.
  - Identify qualitative *categories* of system error and count them.
  - Visualize your embedding spaces with tools like t-SNE or PCA.
  - Visualize your hidden states with tools like LSTMVis.
  - Plot how your model performance varies with amount of data.
  - Ambitious? Build an online demo!

Formative vs. Summative Evaluation

When the cook tastes the soup, that's formative; when the customer tastes the soup, that's summative.

- Formative evaluation: guiding further investigations
  - Typically: lightweight, automatic, intrinsic
  - Compare design option A to option B
  - Tune hyperparameters: smoothing, weighting, learning rate
- Summative evaluation: reporting results
  - Compare your approach to previous approaches
  - Compare different *major* variants of your approach
  - Generally only bother with human or extrinsic evaluations here
- Common mistake: Don't save all your qualitative evaluation for the summative evaluation!

Hyperparameter Tuning

- Tune your baseline
  - You must tune the hyperparameters of your baselines just as thoroughly as you tune them for any new model you propose
  - Failure to do this invalidates your comparisons
  - Don’t tune on the test set!
- Read the fine print while you’re doing your literature review to get a sense of what hyperparameters to worry about and what values to expect.
- If you’re not sure whether to tune a hyperparameter, you probably should.

Hyperparameter Tuning

- Grid search: Inefficient (but common)
- Bayesian optimization: Optimal, but public packages aren't great.
- Good read: random search (Bergstra and Bengio '12) → easy, and near-optimal
  - Define distributions over all your hyperparameters.
  - Sample N times for N experiments.
  - Look for patterns in your results.
  - Adjust the distributions and repeat until you run out of resources or performance stops improving.

Ethics in Data-driven Models

“Instead of relying on algorithms, which we can be accused of manipulating for our benefit, we have turned to machine learning, an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie.”

- Maciej Cegłowski

Example: Pitfalls in Word Embeddings

Occupations most similar to *she*:
nurse, librarian, nanny, stylist, dancer

Occupations most similar to *he*:
architect, captain, philosopher, legend, hero

Impossible to avoid these issues altogether when learning from naturally occurring text.

Source: Bolukbasi et al. ‘16, Quantifyiang and Reducing Stereotypes in Word Embeddings

Data and Biases

- Deploying biased models in the wrong places can lead to harms far worse than bad user experiences
  - Résumé screening, exam scoring, predictive policing...
- Some ML techniques can amplify biases in data
  - Zhao et al. 2017 on multi-label image classifiers:
    - In training data, women appear in cooking scenes 33% more often than men
    - In model's labeling of similar test data, women are detected in cooking scenes 68% more often than men
- Model de-biasing can be complex, political, and maybe even impossible to do fully
  - … and it may harm performance on reasonable metrics.

Data and Exclusion

- Colloquial African-American English isn’t well represented in training data for language identification, parsing, etc.,
  - So technologies like translation and intelligent assistants aren’t as usable for its speakers
  - Result: Users are forced to choose between avoiding their preferred dialect and missing out on the benefits of the technology
- Similar situation for English varieties in Singapore, India, Caribbean, etc., and for regional/minority languages in general

See *Blodgett and O’Connor ‘17*

Talk about the Data

- When discussing your models, be as clear as you can about:
  - What your data looks like, why it was collected, and what kind of information your system learns from it
  - Who (country, region, gender, native language, etc.) produced the text and labels in your dataset
  - Any known biases in your dataset (including the obvious ones)
- Especially important when writing for nontechnical users or clients
- Hard: when possible, build useful confidence metrics
  - Make it clear to the user when the system is out of its comfort zone