Writing and Publishing in NLP

Sam Bowman

(With pieces from Chris Potts and Bill MacCartney's Stanford CS 224U)

Today


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- If you want to publish an influential paper in NLP, the engineering is only half the work.
- This is the last of three lectures about the other half:
  - How to find and understand related work on your problem
  - How to design effective experiments and analyze their results
  - How to stay out of ethical trouble
  - **How to write and publish your work**

Today


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7.—No. 2.

- This is not specific to the class project, but much of it applies.
- For a successful project, you must:
  - Write a clear and carefully-structured paper that is up to the standards of a major conference.

Finding Examples to Work From


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7.—No. 2.

The slides from the *Experimental Design* week covered how to find relevant papers, and how they're structured.

When deciding how to write your paper, use these papers as models for what to include and what to emphasize.

Related: Here are some examples of excellent course papers in this format from past years. (Note the different length limit for 2018.)

Three ways to organize your ideas (Shieber)

- Continental style: "in which one states the solution with as little introduction or motivation as possible, sometimes not even saying what the problem was" [. . .] "Readers will have no clue as to whether you are right or not without incredible efforts in close reading of the paper, but at least they'll think you're a genius."
- Historical style: "a whole history of false starts, wrong attempts, near misses, redefinitions of the problem." [. . .] "This is much better, because a careful reader can probably follow the line of reasoning that the author went through, and use this as motivation. But the reader will probably think you are a bit addle-headed."
- Rational reconstrution: "You don't present the actual history that you went through, but rather an idealized history that perfectly motivates each step in the solution." [. . .] "The goal in pursuing the rational reconstruction style is not to convince the reader that you are brilliant (or addle-headed for that matter) but that **your solution is trivial**. It takes a certain strength of character to take that as one's goal."

Writing the Abstract

- Important for creating a first impression, reviewer bidding, and reviewer assigning.
- A general structure:
  1. The opening is a broad overview — a glimpse at the central problem.
  2. The middle take concepts mentioned in the opening and elaborates upon them, probably by connecting with specific experiments and results from the paper.
  3. The close establishes links between your proposal and broader theoretical concerns, so that the reviewer has fresh in her mind an answer to the question “Does the abstract offer a substantive and original proposal”.

Figure One


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7—No. 2.

It is common and helpful to have the first figure in the paper to explain the main idea of the paper. It should be possible to understand this figure *without* reading the rest of the paper.

Figure One

Abstract

Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network. When trained on the new treebank, this model outperforms all previous methods on several met-

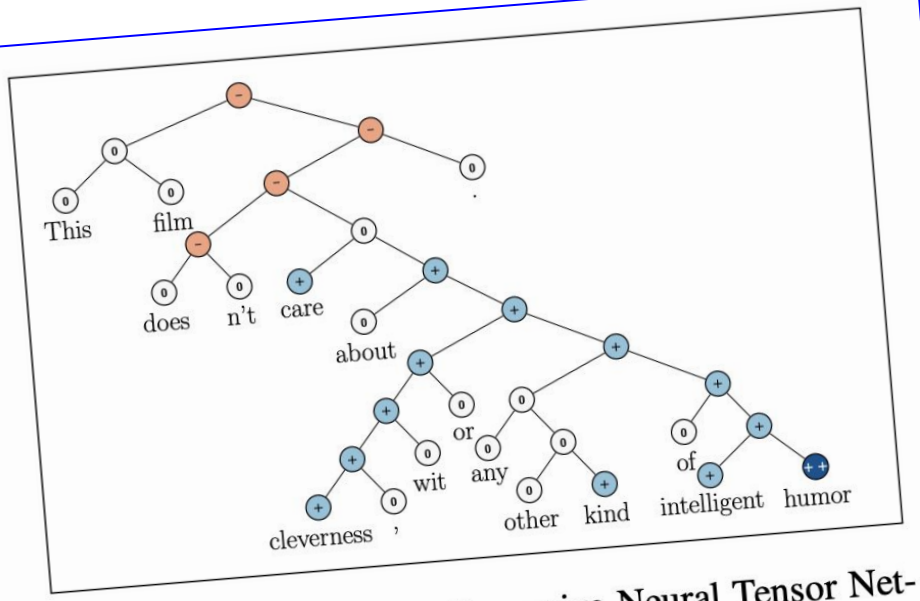


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (− −, −, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

Mistakes to Avoid

Overclaiming


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

The easiest way to get a paper rejected (or given a low grade):

- Saying something that isn't true.

Almost as easy:

- Saying something that's true without sufficient evidence.

LaTeX

Typesetting and LaTeX

- You'll need to use the *LaTeX* typesetting tool for your paper, both for class and for any outside submission.

```
\section{Introduction}

\begin{figure}[t]
    \centering
    \includegraphics[width=\columnwidth]{acl-ijcn
    lp2021-templates/figures/figures_exponential/
    overall.pdf}
    \caption{Overall learning curves for the five
    evaluation methods. For each method, we
    compute overall performance for each RoBERTa
    model tested as the macro average over
    sub-task's performance after normalization.
    We fit an exponential curve which we scale to
    have an initial value of 0 and an asymptote
    at 1. Classifier and MDL probing mainly test
    models' encoding of linguistic features;
    BLiMP tests model's understanding of
    linguistic phenomena; LAMA tests factual
    knowledge; SuperGLUE is a suite of
    conventional NLU tasks.}
    \label{fig:overall}
\end{figure}

Pretrained language models (LMs) like BERT and
RoBERTa have become ubiquitous in NLP. New models
require massive datasets of tens or even hundreds
of billions of words \cite{brown2020gpt3} to
improve on existing models on language
```

When Do You Need Billions of Words of Pretraining Data?

Yian Zhang,[*,1] **Alex Warstadt,**[*,2] **Haau-Sing Li,**[3] **and Samuel R. Bowman**[1,2,3]
[1]Dept. of Computer Science, [2]Dept. of Linguistics, [3]Center for Data Science
New York University
{yian.zhang, warstadt, xl3119, bowman}@nyu.edu

Abstract

NLP is currently dominated by language models like RoBERTa which are pretrained on billions of words. But what exact knowledge or skills do Transformer LMs learn from large-scale pretraining that they cannot learn from less data? We adopt five styles of evaluation—classifier probing, information-theoretic probing, unsupervised relative acceptability judgment, unsupervised language model knowledge probing, and fine-tuning on NLU tasks—and draw learning curves that track the growth of these different measures of model ability with respect to pretraining data volume using the MiniBERTas, a group of RoBERTa models pretrained on 1M, 10M, 100M and 1B words. We find that these LMs require only about 10M to 100M words to learn to reliably encode most syntactic and semantic features we test. They need a much larger quantity of data in order to acquire enough commonsense knowledge and other skills required to master typical downstream NLU tasks. The results suggest that, while the ability to encode linguistic features is almost certainly necessary for language understanding, it is likely that other (unidentified) forms of knowledge are the major drivers of recent improvements in language understanding among large pretrained models.

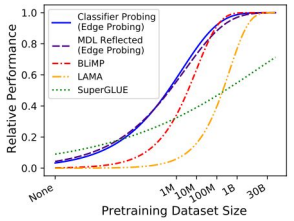
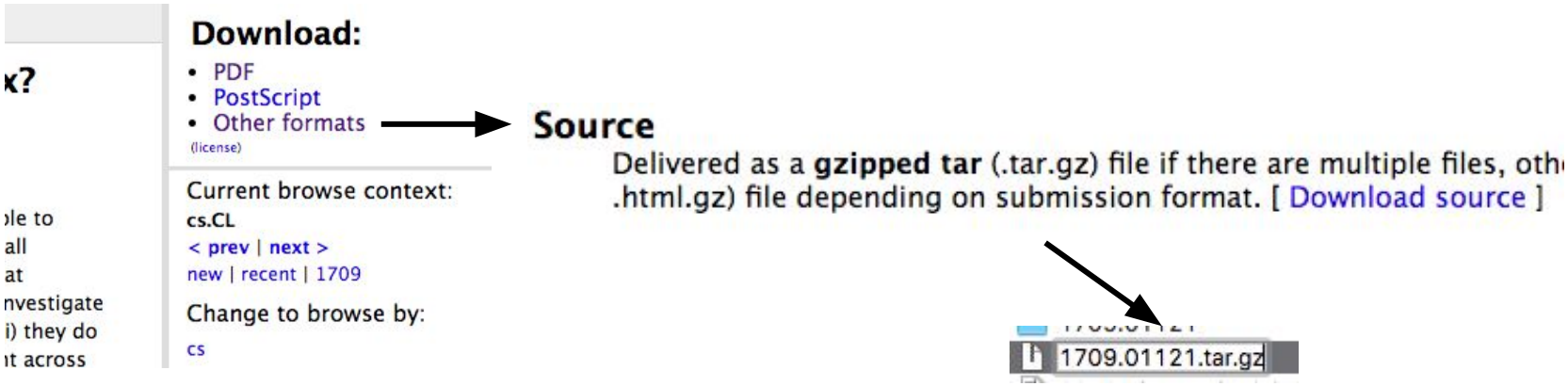


Figure 1: Overall learning curves for the five evaluation methods. For each method, we compute overall performance for each RoBERTa model tested as the macro average over sub-task's performance after normalization. We fit an exponential curve which we scale to have an initial value of 0 and an asymptote at 1. Classifier and MDL probing mainly test models' encoding of linguistic features; BLiMP tests model's understanding of linguistic phenomena; LAMA tests factual knowledge; SuperGLUE is a suite of conventional NLU tasks.

of words), many interesting questions regarding the effect of the amount of pretraining data remain unanswered: What have data-rich models learned

Typesetting and LaTeX

- For the basics of LaTeX, there are good tutorials online.
- We'll hold a lab session with advanced tips that are useful for writing NLP/ML/AI papers.
- Another good source of handy tricks:
  - Look at the LaTeX source code for any paper on arXiv.org



References


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Get comfortable with BibTeX
  - Maintain a .bib file as you do your literature review.
  - Get .bib entries for published papers from most NLP conferences/journals with a simple trick:

    http://aclweb.org/anthology/P/P17/P17-1060.**pdf**

    → http://aclweb.org/anthology/P/P17/P17-1060.**bib**

    (similar tricks sometimes work for other venues, like NeurIPS)
  - Google Scholar distributes .bib entries, but they're **usually wrong**.
  - If a paper has an arXiv version and a published version, you should cite the published version.

Publishing and Presenting your Work

Publishing


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- If successful, finished paper should be the *foundation* of a publishable research paper.
- Major conferences have higher expectations for ambition and for thoroughness of analysis
- Build on what you've found until you have a substantial result that you're confident about, and submit them!

Choosing a venue


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Major conferences:
  - 3–9 page limits (some conferences have multiple options)
  - 3–6 mo review process
  - Low acceptance rate (15–40%) from a pool of well-written papers
- Journals:
  - Review process varies, 2mos–2yrs
    - Often involves one or two chances to improve the paper
  - Viewed as more prestigious by readers outside AI
  - Page limits vary widely

Choosing a venue


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Workshops attached to conferences:
  - Focused on a specific topic, vary year to year
  - ~1mo review process
  - Higher acceptance rate (>50%), lower prestige than main conferences
  - Great for feedback or for a first publication
  - Some don’t formally publish submitted papers, allowing you to submit them to a conference later

Choosing a venue


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- arXiv
  - ~1 day review process: Moderators just make sure the paper fits in the subfield track that it's submitted to.
  - Great way to get feedback/attention for your work:
    - The *majority* of the NLP research community watches arXiv for updates, either directly ('cs.CL' email updates) or through tools like Google Scholar alerts and arxivist.
  - Be careful if you plan to submit your work!
    - Many publication venues impose special rules on papers that are already available online, since they break anonymous peer review.
    - ACL rule, for example: You can only submit arXiv papers to conferences and journals after a one-month waiting period.
  - Readers still expect that the papers are careful and honest. Posting sloppy work will backfire!

Choosing a venue


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Coming up:
  - EMNLP (May 10)
  - CoNLL (June 14)
  - EMNLP workshops (~July–August)
    - WiNLP (August)
  - ICLR (~September–October)
  - Journals: Computational Linguistics/TACL (whenever)
  - arXiv (whenever)

Choosing a venue


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7--No. 2.

- If you're new to publishing in NLP and Machine Learning, *please do not submit to any of these venues* without asking one of us or another more senior researcher.
- For a first paper, it's common to need weeks or months of revisions before your work will be seriously considered.
- Submitting work that isn't ready will waste the time of people who you may later want to ask for reviews/jobs/etc.

Anatomy of an NLP conference submission

Each conference is a little different, but here's a typical process:

1. You submit a completed 8-page* paper, along with area keywords that help determine which committee gets your paper.
2. Reviewers scan a **long** list of titles and abstracts and then bid on which ones they want to do. The title is probably the primary factor in bidding decisions.
3. The program chairs assign reviewers their papers, presumably based in large part on their bids.
4. Reviewers read the papers, write comments, supply ratings.
5. Authors are allowed to respond briefly to the reviews*.
6. The program chair might stimulate discussion among the reviewers about conflicts, the author response, etc. At this stage, all the reviewers see each other's names, which helps contextualize responses and creates some accountability.
7. The program committee does some magic to arrive at the final program based on all of this input.

Anatomy of an NLP conference submission

These rating categories have prose descriptions attached to them to help clarify the program committee's intentions:

Appropriateness:	1-5
Clarity:	1-5
Replicability:	1-5
Originality / Innovativeness:	1-5
Soundness / Correctness:	1-5
Meaningful Comparison:	1-5
Thoroughness:	1-5
Impact of Ideas or Results:	1-5
Impact of Resources:	1-5
Overall Recommendation:	1-5
Presentation Format:	Poster/Talk/Both possible
Best paper possibility?	Yes/Maybe/No

Why papers get rejected


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Overclaiming (common):
  - The paper makes a concrete claim that isn't backed up by appropriate citations or direct evidence.
  - Examples:
    - In an introduction: "Researchers have long struggled to do XYZ." (Do you have evidence that people actually worked on this?)
    - In an introduction: "Unlike current ML models, humans do XYZ by reasoning about concepts like ABC." (Have cognitive scientists really concluded this?)
    - In related work or methods: "BiLSTMs are the best approach to task ABC (XYZ et al. 2018)." (Did XYZ really show that it's *the best* approach? Is this still true now?)
    - In an abstract or conclusion: "We show that our system beats BERT." (Did you run a *fair* comparison with BERT?)

Why papers get rejected


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Overclaiming (common):
  - The paper makes a concrete claim that isn't backed up by appropriate citations or direct evidence.
  - Reviewers *may* allow for some overclaiming when describing related work and background, *if* they trust that you'll fix it. Any overclaiming about your own contributions will result in a rejection.

Why papers get rejected


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Methodological issues (common, related):
  - Your methods are *almost* sufficient to make some interesting claims, but there's a crucial flaw that makes the results hard to interpret.
  - Examples:
    - Did you tune your baseline correctly?
    - If you're working with pretrained models, does the model's tokenization and vocabulary make sense for your task?
    - Did you use the right metric for the claim you're making?

Why papers get rejected


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7.—No. 2.

- Motivation (somewhat common, related):
  - You answer a research question, but don't explain why a reasonable person would ask that question.
  - Examples:
    - If you're using NLP for a problem for the first time, could someone actually use NLP for your problem in the real world without hitting ethical/legal/logistical issues?
    - If you're trying to improve the performance of some system that *isn't* the state of the art, is there some reason that a user would want to use that system?

Why papers get rejected


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Novelty (somewhat common):
  - Did someone else already answer this question? If so, did you explain why it was necessary to revisit the question?
- Impact (less common):
  - Will at least a few dozen people find this paper relevant to their own work in the future?
- Fit (less common for larger conferences, though relevant to workshops):
  - Do members of your intended audience tend to read papers that are published here? Do your scientific peers tend to review for this venue?

Conference presentations


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876
Vol. 7—No. 2.

- So you got in. What's next?
  - Submit a final 'camera ready' version of your paper, incorporating minor changes from the reviews (2-8 weeks).
  - Plan travel to the conference (if there's not an active pandemic).
  - Present your work!

Conference presentations


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Types of presentation:
  - Plenary/general sessions:
    - ~1h talks, usually only for invited senior speakers
    - ~20m talks for 1–6 best papers
  - Parallel sessions (less common in general ML):
    - ~7-20m talks for the top 5–50% of submitted long papers
    - (sometimes) ~3-10m talks for the top 5–50% of submitted short papers
  - Poster sessions:
    - 1.5–4h poster sessions for all other accepted papers

Sharing your code


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Few NLP venues have hard reproducibility requirements, but most have a strong preference for reproducible work.

Sharing your code


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

- Standard practice:
  - Make a GitHub repo for all of your code and saved model files for your best runs.
  - Prepare a readme with instructions on:
    - How to access any necessary data
    - How to install and configure your code
    - How to retrain your model to reproduce your experiments
    - How to run your trained model on new data
    - *All of this should be sufficient to reproduce all of your results without contacting you.*
  - Include a link to the repo in your paper.
    - Best practice is to anonymize the whole repo during blind peer review, but many people just censor the URL and only include it after review.

Your lightning talk

The Lightning Talk

- As part of your project, you'll need to give a *lightning talk* of about three minutes.
  - This format appears at some conferences and workshops, too!
  - (Usually alongside posters.)
- Your goal is to explain your main conclusion to your classmates.
  - Assume that your audience has remembers the key ideas from the lectures/readings for this class, but hasn't read your proposal or any of the prior work on your problem.
  - Make sure your team agrees on what your main conclusion is—you don't have time to talk about more than one point!

Logistics

- You can use up to three slides.
  - You will present from your own computer.
    - Test this in advance! You will not get extra time if your slides don't work.
  - Use any style/layout you want.
- Not everyone on the team needs to speak.
  - But everyone on the team should help develop the presentation.
- You have two minutes to answer an audience question.
  - *Everyone* on the team should be able to answer *any* question.
    - This is a fundamental part of coauthorship on academic work: By agreeing to be an author on some piece of academic work, you are claiming that the entire work is correct. That means that you need to understand the entire work well, even if parts of it are outside your normal area.

My lightning talk (example)

A large annotated corpus for learning natural language inference

Sam Bowman, Gabor Angeli, Chris Potts, and Chris Manning (EMNLP 2015)

If...

Two dogs are running through a field.

Is it true that...

Some puppies are running to catch a stick.

Maybe.

A large annotated corpus for learning natural language inference
Sam Bowman, Gabor Angeli, Chris Potts, and Chris Manning (EMNLP 2015)

The largest dataset for this task has only 5000 training examples, and many people have tried and failed to train neural networks on the task (Marelli et al. 14).

A large annotated corpus for learning natural language inference
Sam Bowman, Gabor Angeli, Chris Potts, and Chris Manning (EMNLP 2015)

The results are reasonable, if a bit simple and repetitive:

S1: A soccer game with multiple males playing.
S2: Some men are playing a sport.
Is S2 true? *Yes.*

Your lightning talk

Tips

- Don’t put anything on your slides that you won’t talk about.
- Don’t talk faster than you do normally.
  - If the listener wants more information, they can ask questions...
  - ...but if they can’t keep up with the pace of the talk, they’ll just stop paying attention.
- Be honest about your conclusions and limitations.
- It’s okay to present a slightly simplified version of your idea, as long as you’re not misleading the audience.
  - Only use technical terms if you have time to explain them!
- Practice, simplify, and practice again!
  - For a short talk, 25+ rehearsals is normal.

More resources

More resources and advice


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

Some advice from NYU researchers:

- From someone who's won awards for writing:
  - "I realize this is frustrating advice and probably not what they want, but I would recommend [just reading] a lot of stuff, focusing on nontechnical writing written by [fluent] English speakers. So like, not arXiv. The Atlantic[?]"
- Get feedback from the NYU Writing Center a few times while you're here, especially when working in a style that's new to you.
- Classic writing about writing:
  - Style: Lessons in Clarity and Grace
  - Clear Technical Writing
  - Politics and the English Language

More resources and advice


The California Agriculturist
AND
LIVE STOCK JOURNAL.
SAN JOSE, CAL., FEBRUARY, 1876

Some advice from NYU researchers:

- Most philosophy classes will emphasize the kind of precise writing you'll need for technical papers.
- Prominent NLPers on how they spend the week before a paper deadline
- A nice math style guide for NLP
- More ML writing advice

Next Steps

- You should have baseline code ready very soon (if you don't already).
  - Debugging experimental code usually takes much longer than writing it. For some advice on that, see Graham Neubig's slides here.
- Start running experiments as soon as you can.
  - The Greene cluster should have plenty of capacity now, but it always gets more competitive near the end of the term.
  - Unless the cluster goes down for multiple days, we won't give extensions because of resource access issues.

Coming Up

- Tomorrow:
  - Lab/lecture on a typical NLP experimental codebase with Nikita
  - Jacob Devlin on BERT at NLP/Text as Data series (4p, CDS)
- Draft paper due in four weeks
- Next week:
  - Semantics crash course