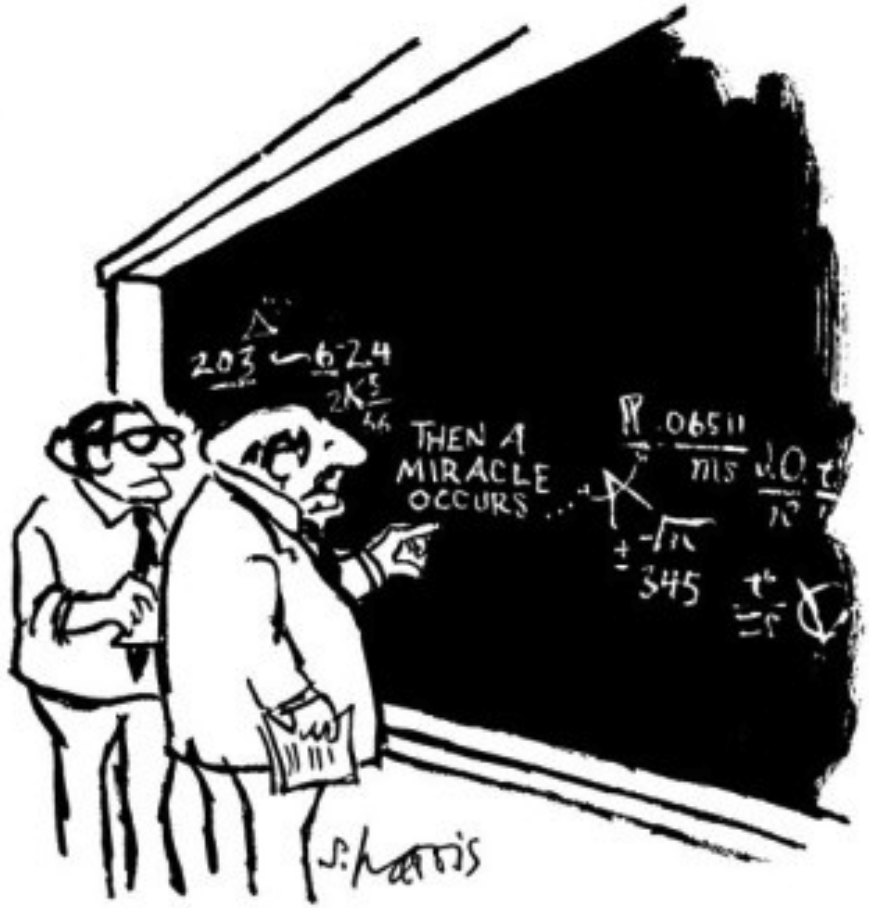
Wrap-up

Bill MacCartney and Christopher Potts

CS 244U: Natural language understanding
Mar 7

Experiments in NLP



"I think you should be more explicit here in step two."

http://www.condenaststore.com/-sp/I-think-you-should-be-more-explicit-here-in-step-two-Cartoon-Prints_i8562937_.htm

Training/Development/Test splits

1. Your final experiments should be done on a **test** set that was not used at all during development.
2. Thus, it is imperative that your model handle new data well.
3. Thus, you should divide your non-test data into a **training** set and a **development** set.
4. You'll want to do lots of testing of features and different hyperparameters during your research.
5. For this phase, you should consider experiments involving cross-validation **on your training set only**.
6. Use the development set only sparingly; you don't actually want to optimize your model to that data, since it differs from your test data.

Benchmarks

1. *Weak baselines*: random, most frequent class
2. *Strong baselines* (and the desirability thereof): existing models and/or models that have a good chance of doing well on your data
3. *Upper bounds*: oracle experiments, human agreement (non-trivial; human performance is rarely 100%!)

Comparisons with other approaches

- Where there are published results for exactly the data you are working with, this is pretty straightforward. In such situations, comparisons are essential — your paper will likely be rejected without them.
- Where there are published results on different but related data, or where your goals differ slightly from those of published authors, comparison is equally important but much trickier. You have to make a plausible case for both the comparisons and your approach.
- Where there are no related published results, the comparisons won't be quantitative, but rather conceptual. In that case, they should appear earlier in your paper, to help readers conceptualize what you are doing.

Evaluation contexts

- Intrinsic evaluation: how a system performs relative to its own objective function
- Extrinsic evaluation: how a system contributes to a larger more complex task or system.

Evaluation techniques

Understanding your system's performance:

- Confusion matrices to spot problem areas and overlooked oddities.
- Visualization to make multiple formal and informal comparisons and identify overlooked relationships.

Evaluation metrics

- Accuracy: appropriate only for balanced datasets, and where all the categories are of equal value to you.
- By-category precision and recall: measurements that abstract away from category size and, together, help avoid rewarding conceptually poor behavior.
- F1: harmonic mean of precision and recall, appropriate only where both are of equal importance and are at least roughly comparable. Use weighted variants to favor precision or recall.
- Macroaveraged F1: average of F1 scores for the classes. Equal weight to each class.
- Microaveraged F1: pools the by-category contingency tables into a single one and then computes F1. Mostly controlled by the largest classes.
- Rank correlation: for ordered predictions
- Specialized metrics: some fields have their own preferred evaluation techniques; it is essential to be aware of such norms.

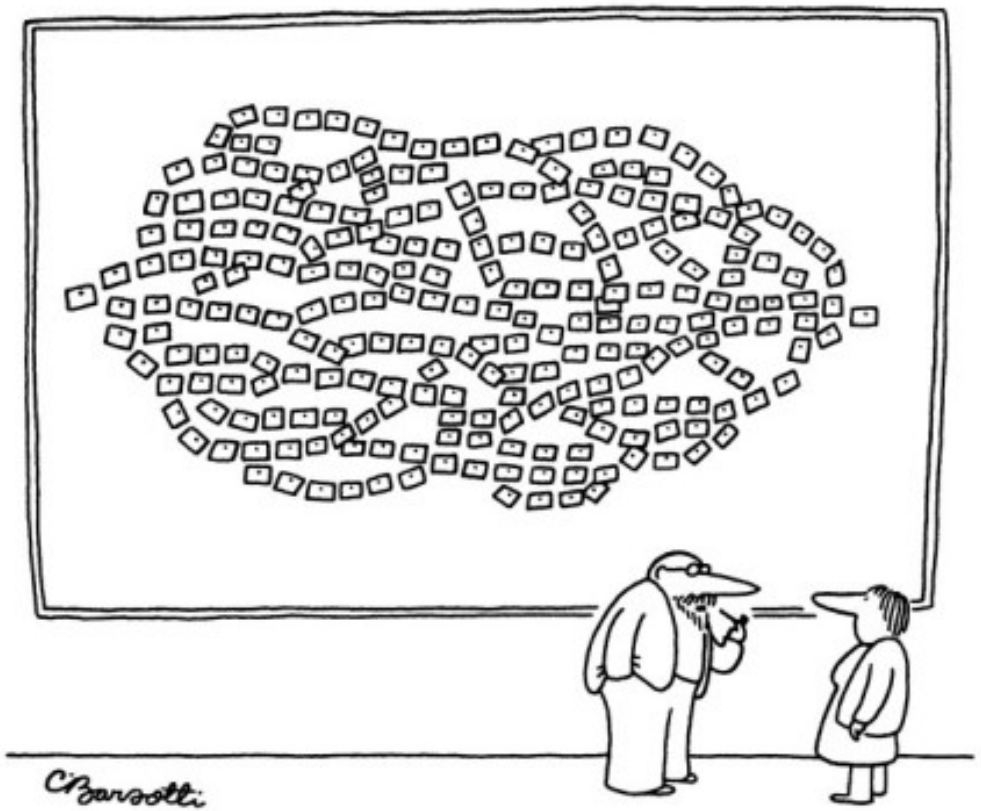
Things to investigate and report on

- Feature ablation/accretion studies
- Learning curves
- Evaluations on different datasets
- Time and space requirements
- Significance testing the easy way, with approximate randomization: `http://masanjin.net/sigtest.pdf`

Development methodology

1. Construct a tiny toy data set for use in system development
2. Iterative development:
   a. Get a baseline system running on real data ASAP.
   b. Implement an evaluation — ideally, an automatic one, but could be more informal if necessary.
   c. Hill-climb on your objective function, using human intelligence.
   d. Feature engineering cycle: add features ⇒ eval on development data ⇒ error analysis ⇒ generalizations about errors ⇒ brainstorming ⇒ add features
3. Research as an "anytime" algorithm: have some results to show at every stage
4. Consider devising multiple, complementary models and combining their results (via max/min/mean/sum, voting, meta-classifier, . . .).
5. Grid search in parameter space:
   - can be useful when parameters are few and train+test is fast
   - easy to implement
   - informal machine learning

On writing papers



"It's plotted out. I just have to write it."

http://www.condenaststore.com/-sp/
It-s-plotted-out-I-just-have-to-write-it-New-Yorker-Cartoon-Prints_i8542726_.htm

A commonly-used structure for NLP papers

1. Opening: general problem area, goals, and context.
2. Related work (if it helps with set-up; else move to slot 6)
3. Model/proposal
   a. Data (separate section if detailed/new/...)
   b. Experimental set-up
4. Results
5. Discussion
6. Related work (if here largely for due diligence, or if understandable only after the results have been presented)
7. Conclusion: future work — not what you will do per se, but rather what would be enlightening and important to do next.

Similar to the format for experimental papers in psychology and linguistics, except that they tend to have much longer openings and section 3 often has more sub-parts on the methods used.

Stuart Shieber on the 'rational reconstruction' format

Full note:
`http://www.stanford.edu/class/cs224u/slides/schieber-writing.pdf`

- Continential 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."

David Goss's hints on mathematical style

"Two basic rules are: 1. Have mercy on the reader, and, 2. Have mercy on the editor/publisher. We will illustrate these as we move along."

`http://www.math.osu.edu/~goss.3/hint.pdf`

On conference submissions

http://xkcd.com/541/

Typical NLP conference set-up

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.

Typical ACL set-up

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
Resubmission as short paper:	recommended/not recommended

Presentation types and venues

Presentation types

- Oral presentations vs. poster presentations
- Workshops vs. main conferences

Some important NLP conferences (broadly construed)

- ACL
- NAACL
- EACL
- COLING
- EMNLP
- CoNLL
- ICWSM
- AAAI
- ICML
- NIPS
- CogSci

Typical linguistics/cog-sci set-up

1. You submit an abstract or short form paper.
2. The reviewers write comments and give rankings.
3. The program committee does some magic to arrive at the final program based on this input.

On abstracts

- 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”.

On giving talks

http://www.condenaststore.com/-sp/Sign-GONE-LECTURIN-New-Yorker-Cartoon-Prints_i8476260_.htm

Basic structure

- Beginning
  - What problem are you solving?
  - Why is it important?
  - What approaches have been tried, and why have they not fully solved the problem?
- Middle
  - What data?
  - What approach? (model type, feature representations)
  - hHow to evaluate success?
- End
  - Quantitative results, graphs that slope upward.
  - Qhich features/techniques/ resources contributed most?
  - What kinds of things do we still get wrong? Examples.

(Mirrors paper stucture, but talk structure has to be simpler.)

Pullum's Golden Rules

Geoff Pullum's Five Golden Rules (well, actually six) for giving academic presentations

1. Don't ever begin with an apology.
2. Don't ever underestimate the audience's intelligence.
3. Respect the time limits.
4. Don't survey the whole damn field.
5. Remember that you're an advocate, not the defendant.
6. Expect questions that will floor you.

`http://www.lel.ed.ac.uk/~gpullum/goldenrules.html`

Honesty

Patrick Blackburn's fundamental insight:

Where do good talks come from?

Honesty.

"A good talk should never stray far from simple, honest communication."

Slide contents: two schools of thought

Minimalism

1. Your slides should be as spare as possible without sacrificing clarity.
2. The audience should spend most of the time listening to and looking at you.
3. Individual slides do not stay up for long or get used in more than one way.

Comparative

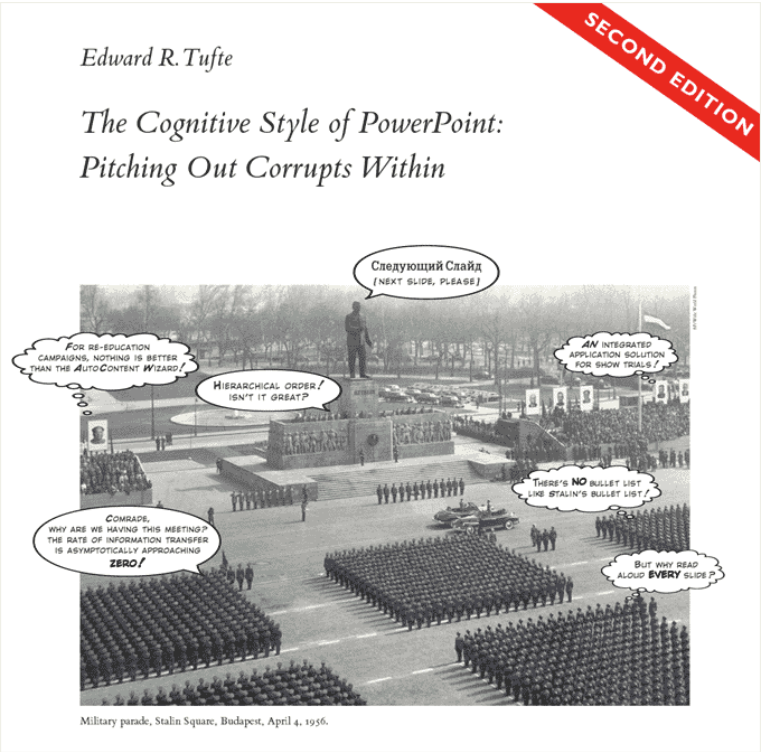
1. Your slides should be as full as possible without sacrificing clarity.
2. Your talk should make it easy for people to spend time studying your slides.
3. Individual slides stay up for a long time and get used to make multiple comparisions and establish numerous connections.

Slide contents: two schools of thought

A personal matter

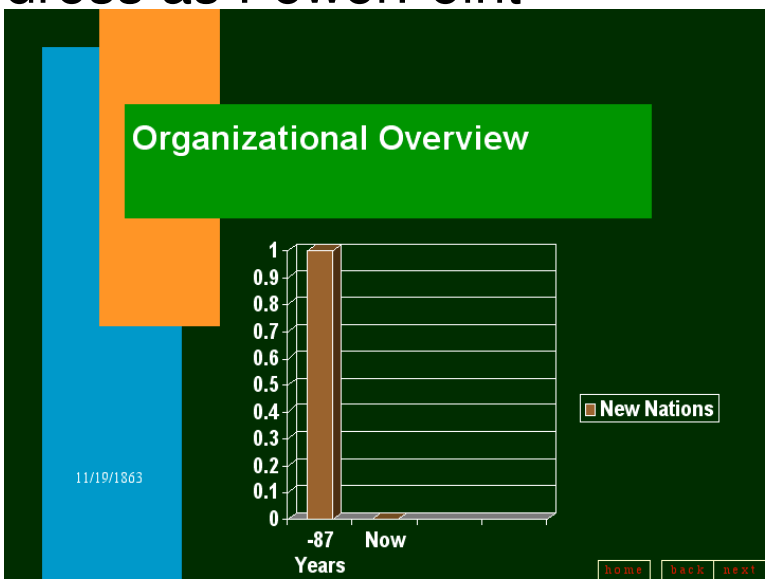
- The minimalist view seems right for telling a story — often the best mode when time is of the essence and the audience is mainly there to learn about what your paper contains.
- The comparative view seems right for teaching; it's the closest slides come to a full, well-organized chalkboard.
- Find the style that works for you. As long as you think long and hard about what it will be like to listen to your talk, and adjust accordingly, you'll shine.

PowerPoint is evil anyway



http://www.edwardtufte.com/tufte/powerpoint

Peter Norvig: Gettysburg Address as PowerPoint



http://norvig.com/Gettysburg/

More mundane things

- Turn off any notifications that might appear on the screen.
- Make sure your computer is out of power-saver mode so that the screen doesn't shut off while you are talking.
- Shut down running applications that might get in your way.
- Make sure your desktop is clear of files and notes that you wouldn't want the world to see.
- If using PowerPoint: have a PDF back-up just in case.
- Projectors fail often; always be prepared to give the talk without slides.

The question period

- This is the most important part of the conference presentation.
- It should be a chance for the audience to gain a deeper understanding of your ideas. When the entire question period has this aim, it is a joy.
- But sometimes other things happen.
- Try to pause for one second before answering each question.
- Never say “I have no idea” and leave it at that.
- When floored, say: “I have no idea, but what . . . ”
- Most questions wont make total sense to you. Your questioner doesnt know the work all that well.
- You’ll be a hit if you can warp every question you get into one that makes sense and leaves everyone with the impression that the questioner raised an important issue.

Your presentations

The Night Before the Big Meeting Frank Receives a Visit from the PowerPoint Fairy.

http://www.condenaststore.com/-sp/The-Night-Before-the-Big-Meeting-Frank-Receives-a-Visit-from-the-PowerPoin-New-Yorker-Cartoon-Prints_i8544502_.htm

Lightening talk limitations

- You have only 4 minutes!
- Prepare — you have only 4 minutes, and thus you can't waste time repeating yourself, figuring out how you want to state things and so forth.
- Practice — nothing in your slides should surprise you; for every slide, you should have a rhetorical plan of action.
- Coordinate — if you more than one person from your group is speaking, practice the transitions carefully so that they don't waste time.

Practical details

- Give your presentation file an informative name — “talk.pdf” is bad.
- Send us your slides in advance if possible.
- Otherwise, bring them on a thumb drive.
- Don’t rely on the presentation machine having specific fonts. (PDF is safest)
- When you are on deck, wait against the wall near the projector.