Announcements

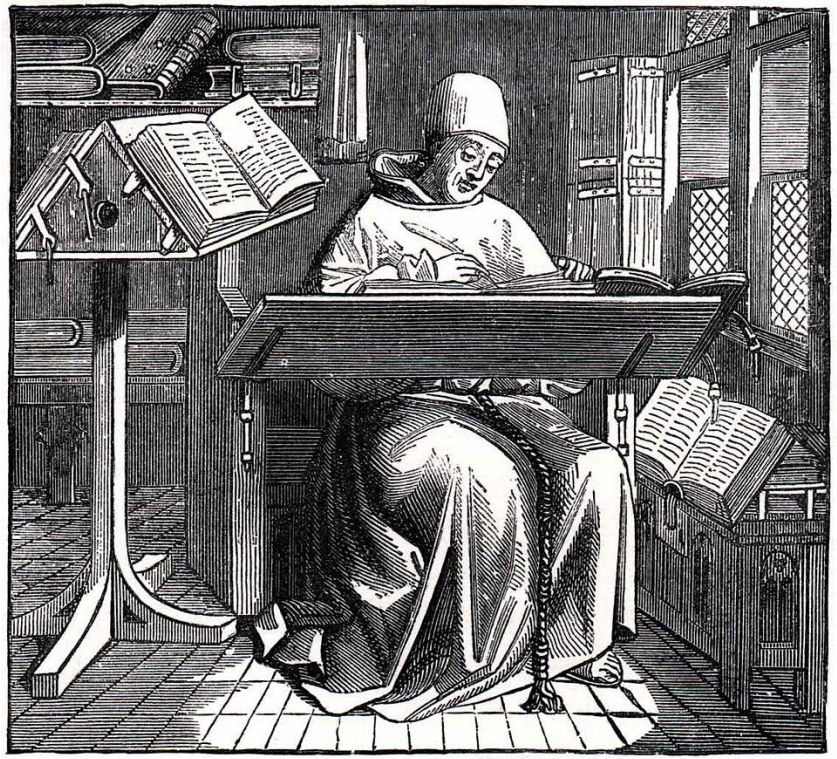
- Project proposals are due today, 11:59pm
- We will assign one TA per team thereafter
- Thursday classes will be split up into five 15min slots
- We'll send out a sign-up form for slots tomorrow

Writing well (in ML/AI)

CSE 481M

Pang Wei Koh's opinions

(slides adapted from Noah Smith and Chris Dyer)



SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

Writing is a skill

- You will get better by **practicing** it
- You will get better by getting feedback
- You will get better by reading good writing!

- Not a native English speaker?
  - Not a problem!
  - Good research writing is about **good ideas** and **clear thinking**, not a big mental lexicon

Tips for writing a paper

Your job as a writer

- You are writing for **your readers**
  - To teach your reader something you figured out
  - To convince your reader of something
- Therefore, your job:
  - Is **not** to show how clever you are
  - Instead, your readers should feel clever as they learned something new

Next we derive an analytic expression for the error estimate confirming the empirical findings. Obviously

$$\Sigma_n[f] = \sum_{k=1}^{n} w_k f(x_k) = \frac{1}{2\pi i} \int_C d\zeta \, \frac{1}{\zeta} \left(1 - \frac{R_{2n-1}(\pi\zeta)}{S_{2n-1}(\pi\zeta)}\right) \frac{1}{\zeta^2} f\left(\frac{1}{\zeta^2}\right), \tag{4.2}$$

https://www.reddit.com/r/mathmemes/comments/1bgl2ta/proof_by_obvious/

Who is your reader?

- For this class (and for most technical papers), your audience:
  - Researchers who are experts in your problem
  - Researchers working on broadly related problems

- Also:
  - Researchers today
  - Researchers N years in the future

https://www.reddit.com/r/timetravel/comments/orj5y5/time_travel_meme/

Imagine your reader

- Knowing your reader lets you determine:
  - What notation they are familiar with
  - What terminology is appropriate
  - What level of detail is appropriate
- **Respect** your reader
  - Don’t overclaim!
  - Get to the point
  - Don’t make the reader work more than necessary
  - Establish common ground
  - Don’t be too harsh

Pitfalls when imagining your reader

- Do not overestimate your readers
  - We are not as knowledgeable as you!
  - We are not as clever as you!
  - We have not read everything you had read, or as carefully!
  - We will read your paper in minutes, hours, or days…
    you have worked on it for weeks, months, or years!

Questions to ask (when writing for your reader)

- Are you introducing a new problem?
  - Is the problem obviously important or do you need to convince them it's important?
- Are you introducing a new technique?
  - What's been tried previously?
  - Benefits and costs relative to alternative techniques? (be honest)
- What is difficult to understand?
  - Algorithms [correctness, complexity]
  - Theorems [proofs, intuitions]
  - Models [assumptions]
  - Experimental setup [what questions are answered]

A typical conference paper structure

- Title (1000 readers)
- Abstract (1 paragraph, 100 readers)
- Introduction (1 page, 40 readers)
- The problem (1 page, 10 readers)
- The idea (2 pages, 10 readers)
- The details (3 pages + more in appendix, 3 readers)
- Related work (1 page, 10 readers)
- Conclusions and future work (0.5 pages)

Structuring a paper

- Start with the known, move to the new
- Starting out:
  - Identify a problem in need of solving
  - Identify an example illustrating some unexplained phenomenon
- Progress logically to new material
  - What is your proposed solution/explanation? What's the key idea?
  - How is it related to past work?
  - Why did you choose this solution?
  - How do you express your solution formally?
  - What did you do to realize this solution?
  - Results, analysis

Structuring a paper

- What is logical structure?
  - Getting you to the idea / insight / contribution in the most direct way
  - Maximizing information content

- What is **not** logical structure?
  - Recapitulating how you got to an idea – don't make your reader suffer the way you did!
  - Giving unnecessary or irrelevant background
  - Focusing on unimportant aspects
  - Being overly defensive

The introduction

- Identify the problem you are solving
- Clearly describe your contributions
  - These drive the structure of the whole paper

- The first sentence should only make sense for your paper, and not for any random paper in the same field
  - No: “LLMs have the potential to revolutionize the world but they carry risks in high-stakes applications…”
  - Yes: “An ideal model evaluation ought to (1) identify where the evaluated model fails in a human-interpretable way, and (2) provide actionable guidance to improve the model…”

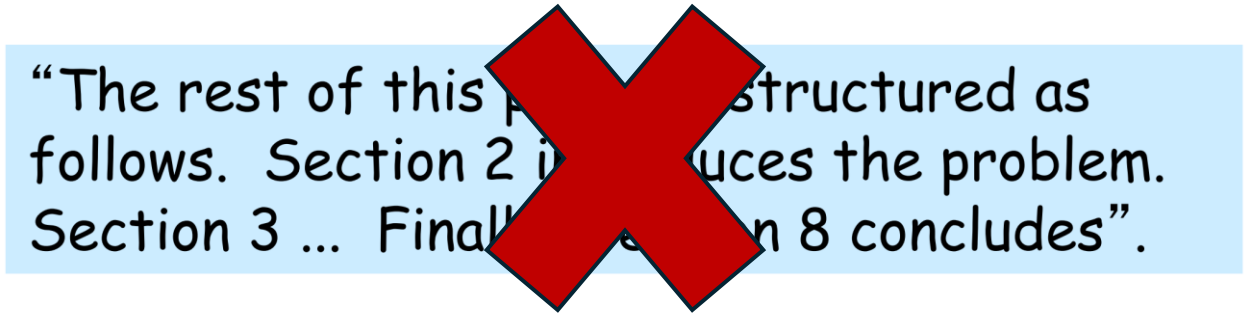
The introduction

- Use a running example throughout the paper
- Pick examples that:
  - Illustrate the easy case easily
  - Illustrate the simplest complicated case easily
  - Are concrete and real
- Concrete → abstract

Your idea

- Figure out what your key idea is
- Make sure the reader knows what your idea is. Be 100% explicit!
  - “The main question we ask in this paper is…”
  - “The key idea behind our method is…”
- This belongs at the very beginning of the paper
  - The research should be surprising, not the writing
- Good ideas that are not distilled = bad paper

No “the rest of this paper is…”



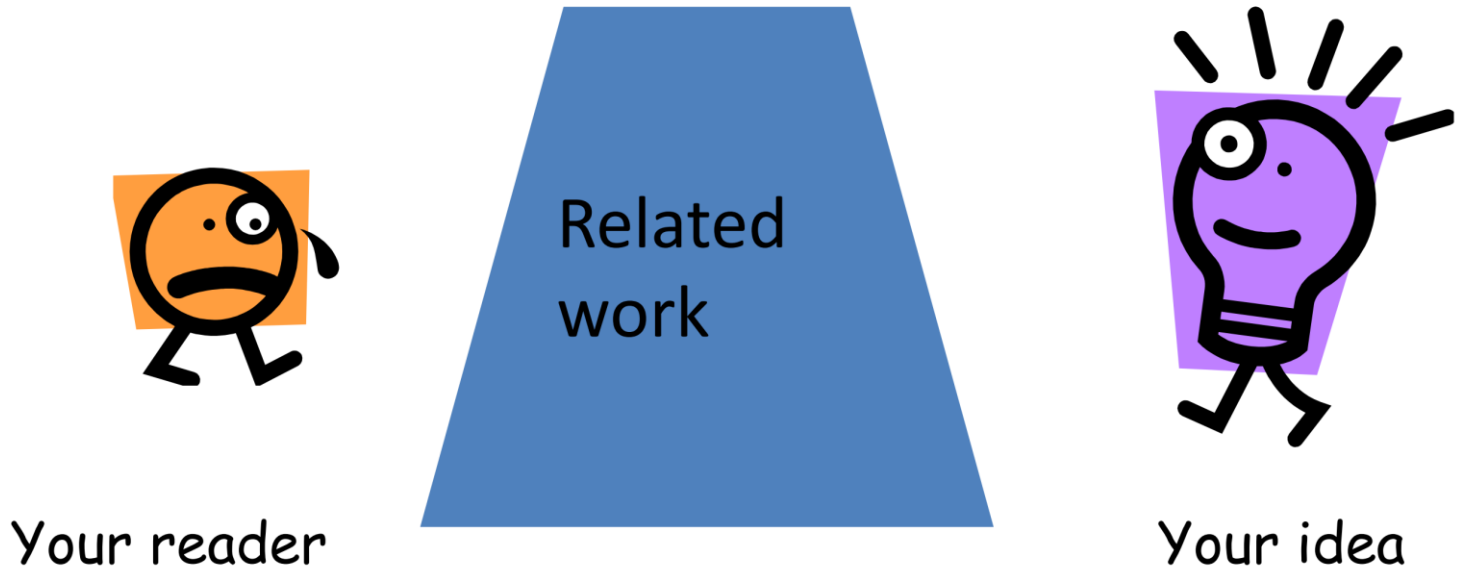
- Instead, use forward references from the narrative in the intro
- The intro should give a road map of the whole paper, and therefore forward reference every important part

In this work, we solve these problems using a Monte Carlo technique with none of the above drawbacks. Our technique is based on a novel Gibbs sampler that draws samples from the posterior distribution of a phrase-based translation model (Koehn et al., 2003) but operates in linear time with respect to the number of input words (Section 2). We show that it is effective for both decoding (Section 3) and minimum risk training (Section 4).

A typical conference paper structure

- Title
- Abstract
- Introduction
- ~~Related work~~
- The problem
- The idea
- The details
- Related work
- Conclusions and future work

No related work yet!



We adopt the notion of transaction from Brown [1], as modified for distributed systems by White [2], using the four-phase interpolation algorithm of Green [3]. Our work differs from White in our advanced revocation protocol, which deals with the case of priority inversion as described by Yellow [4].

No related work yet!

- Problem 1: the reader knows little about the problem yet; so your (carefully trimmed) description of various technical tradeoffs is incomprehensible

- Problem 2: describing alternative approaches gets between the reader and your idea

How to write about related work

Make your laundry list/annotated bibliography version cites all the things. Tuck it at the end of the paper.

As you write, move citations from the laundry list into the paper.

The most important papers your paper is "conversing with" go in the introduction.

Papers that are part of your narrative should be smoothed in where they fit naturally (problem, idea, details, …).

Finally, smooth the "unmatched socks" into a coherent, organized related work section that discusses more distant works and larger context, also potential confusions.

Dyer et al. (2013) use similar terminology to refer to a different idea in a different context …

Tips for writing in general

Get started

- Writing is the best way to develop and clarify your ideas
- You may not have a completely focused idea when you **start**, but you **must** have a completely focused idea when you **finish**

Ask people for help

Ask as many colleagues/friends as possible to read your work!

- Explain what you want (“I got lost here” is much more important than “bayes should be capitalized”.)
- Ask your reader to explain your contribution back to you. Did they get it right?
- An expert can check details, but the logic of any paper should be comprehensible to a non-expert.
- Each reader can only read your paper for the first time once!

Use simple, direct language

NO	YES
The object under study was displaced horizontally	The ball moved sideways
On an annual basis	Yearly
Endeavour to ascertain	Find out
It could be considered that the speed of storage reclamation left something to be desired	The garbage collector was really slow

To hedge or not to hedge?

- Empirical science is about failing to refute an idea, not about proving that an idea is correct.
  - Your language around conclusions should signal your awareness of this, e.g., “our results suggest…” vs. “we proved that…”
- Don’t hedge on established facts!
  - Leave your beliefs out of it; focus instead on the reasons for those beliefs.
  - Watch out for verbs like believe and seem.
  - “The results ~~suggest that~~ show that X goes up as Y goes down”
- Clearly differentiate between hypotheses and observations
  - “We conjecture that…”, “One potential explanation is that…”

Terminology

- Developing good nomenclature is hard
- Adhere to community expectations; don't try to invent totally new nomenclature to replace what readers already know.
- Good nomenclature is honest and transparent.
  - Bad: "conditional random fields," "transformer"
  - Better: "finetuning," "word vector," "recurrent neural network"
- Avoid anthropomorphization

Verbs

- Use strong verbs
  - “We introduce the GAGA algorithm” > “We propose the GAGA algorithm.”
  - Good verbs: introduce, validate, verify, demonstrate, show, prove
- Use unambiguous verbs
  - “We now present the wombat feature...” Did you invent it? Are you reviewing it? Present is ambiguous. Use an unambiguous verb!
- Clearly convey the correct degree of belief
  - Show, demonstrate, suggest, hypothesize, conjecture

Adjectives

- Avoid value-judgment adjectives.
- Bad:
  - We introduce an important algorithm.
- Less bad [verifiably true]:
  - We introduce a novel algorithm.
- Better [true and precise]:
  - We introduce a novel, polynomial time decoding algorithm using a linear program relaxation of the ILP.

Nouns

- This
  - Try to avoid pronoun this “This raises questions…”.
  - Prefer instead demonstrative this: “This pattern of results raises questions…”
- Citep vs citet
  - (Smith et al., 2012) is not a noun.
  - However, *Smith et al. (2012) offered an intriguing solution to the problem of nouns.*

Discourse connectives

- The end of every sentence is an opportunity for a reader to get bored and give up.
- Discourse connectives signal the logical relationship that the next sentence will have to what came before. This keeps them going:
  - However,
  - As a result,
  - Therefore,
  - On the other hand,

Math

- Math formalism can help you clearly state what you mean
- Good math should clarify and not obscure
- Use math if and only if it clarifies your message

- Use words liberally:
  - Remind the reader what symbols mean
  - Explain the intuition behind the math

Polish

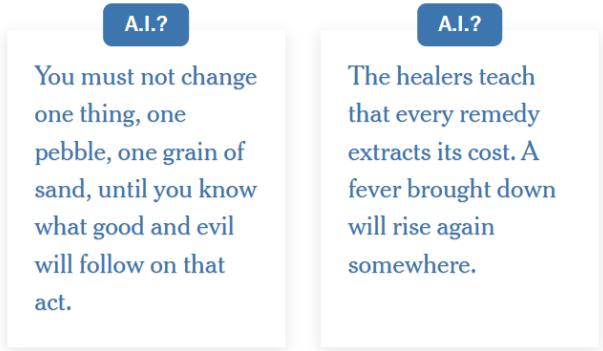
- There are hundreds of little conventions good writers follow, often compulsively, such as:
  - Spelling, punctuation, grammar norms, “e.g.,” vs. “i.e.,” vs. “cf.”
  - Citation styles
  - Use of italics, boldface, abbreviations, …
  - Managing tables and figures: self-contained, clear captions; references in the main text; ease of reading; font size; color-blind-friendly palettes, …
- Making these things perfect will not save an unclear paper!
- But a lack of polish will distract readers from your ideas and make it harder for them to trust you.
- Be the kind of author/scientist who pays attention to details!

Your voice

- True findings are true no matter who found them; we write with some personal distance from the content, and this establishes trust.
  - No: happily, our method worked better than the baseline
  - Informal language and slang deplete reader trust

- Readers suffer if all papers sound the same.
  - Avoid clichés, tropes, catch-phrases, repetition, dry writing without variation, …
  - Avoid sentences that many other people/papers could also have written
  - Be professional but also engaging
  - Proofread by reading your paper out loud.
  - Pay attention to other writers' style when you read!

What about LMs?



Who's a Better Writer: A.I. or Humans? Take Our Quiz.

What about LMs?

- LMs can be great writing partners and beta readers
- Treat them like one!
  - Feedback is valuable…
  - But take what they say with a heap of salt
  - Ask them for feedback, not to write it for you!
- You'll learn to write and develop your own voice by practicing

- Aside: As of today, LM-generated text is still fairly obvious

Summary

- If you remember nothing else from today:
  - Write for your readers, not yourself
  - Identify your contributions and the key ideas
  - Use clear, concrete examples, then move to the abstract
  - Use precise language, cut the fluff
  - Use LMs like you would a (human) writing partner
  - Writing is a skill you can practice!

Thanks to & further material

- Noah Smith and Chris Dyer's slides
- Philip Resnik (UMD)
- Simon Peyton Jones (MSR Cambridge)
- Jason Eisner (JHU)
  - http://www.cs.jhu.edu/~jason/advice/how-to-write-a-thesis.html
- Geoffrey K. Pullum (Edinburgh)
  - http://www.lel.ed.ac.uk/grammar/passives.html