Announcements

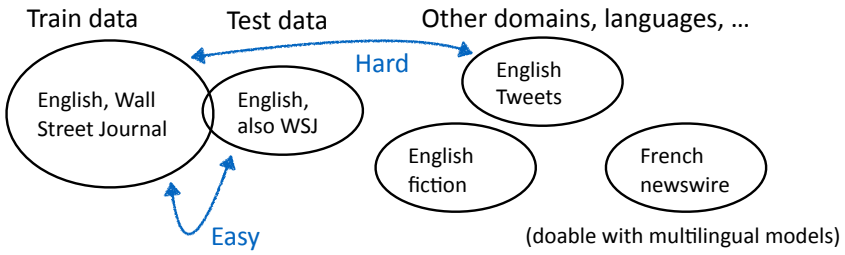
- A5 due today
- Final project released (more details at the end of today's lecture)

Recap

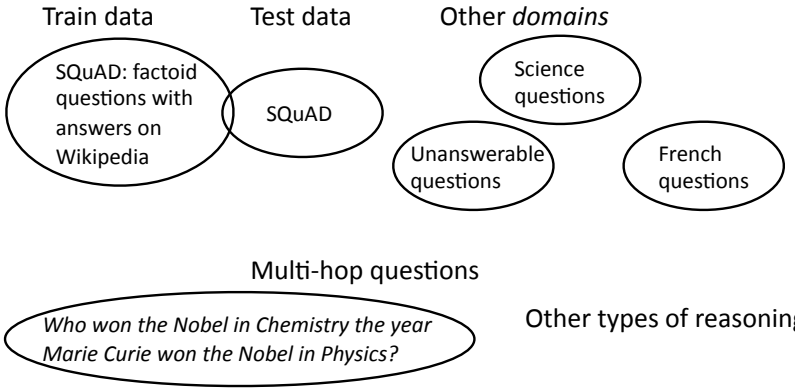
- **Pretraining** (BERT):
  - Train a big model to fill in masked-out words, then adapt it to other tasks. Led to big gains in **question answering** and **NLI** performance:
- **Question answering** (QA):
  - "What was Marie Curie the first female recipient of?"
    -> "The Nobel Prize" (assuming your context document contains the answer)
- **Natural language inference** (NLI):
  - "But I thought you'd sworn off coffee."
    contradicts "I thought that you vowed to drink more coffee."

Generalization

- When a model does well on training data but poorly on test data, we say it *doesn't generalize*
- Many notions of generalization. Example: POS tagging



Generalization: QA



Generalization

- Just doing well on a single test set **is not that useful**
- We want POS taggers, QA systems, and more that can generalize to new settings so we can deploy them in practice
- Sometimes, you can get **very good test performance** while training a **very bad model**. How does this happen?

Annotation Artifacts, Reasoning Shortcuts

Annotation Artifacts

- Some datasets might be easy because of how they're constructed, especially in QA and NLI

What becomes of Macbeth?

What happens to Macbeth at the end?

What does Macduff do to Macbeth?

What violent act does Macduff perform upon Macbeth?

- All questions have the same answer. But some are more easily guessable

QA: Answer Type Heuristics

- Question type is powerful indicator. Only a couple of locations in this context!

QID: 57c5914570fc4995b2b9daa3e5dffb83
Question: where did luther spend his career ?
Answer: university of wittenberg

Start Distribution

```
on 19 october 1512 , he was awarded his doctor of theology and , on 21
october 1512 , was received into the senate of the theological faculty of
the university of wittenberg , having been called to the position of
doctor in bible . he spent the rest of his career in this position at the
university of wittenberg .
```

- Even in more complex settings, can often find plausible answers with a short prefix of the question ("*which president*", "*what violent act*"…)

NLI: Hypothesis-only Baselines

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- What's different about this neutral sentence?
  - To create neutral sentences: annotators *add information*
- What's different about this contradictory sentence?
  - To create contradictions: annotators *add negation*
- These are not broadly representative of what can happen in other settings. There is no "natural" distribution of NLI, but this is still very restrictive

NLI: Hypothesis-only Baselines

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

- Models can detect new information or negation easily
- Models can do very well *without looking at the premise*

Performance of models that only look at the hypothesis: ~70% on 3-class SNLI dataset

	Hyp-only model	Majority class	
SNLI	69.17	33.82	+35.35
MNLI-1	55.52	35.45	+20.07
MNLI-2	55.18	35.22	+19.96

Gururangan et al. (2018); Poliak et al. (2018)

NLI: Heuristics

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

- Word overlap supersedes actual reasoning in these cases
- They create a test set (HANS) consisting of cases where heuristics like word overlap are misleading. Very low performance

McCoy et al. (2019)

Contrast Sets

- How do we control for annotation artifacts? Things like "premises and hypotheses overlap too much" aren't easy to see!
- For any particular effect like lexical overlap, we could try to annotate data that "breaks" that effect
- Issue: breaking one correlation may just result in another one surfacing. How do we "break" them all at the same time?
- Solution: construct new examples through *minimal edits that change the label.*

Gardner et al. (2020)

Contrast Sets

Hardly one to be faulted for his ambition or his vision, it is genuinely unexpected, then, to see all Park's effort add up to so very little. . . . The premise is promising, gags are copious and offbeat humour abounds but it all fails miserably to create any meaningful connection with the audience.
(Label: Negative)

Hardly one to be faulted for his ambition or his vision, **here we see all Park's effort come to fruition.** . . . The premise is **perfect**, gags are **hilarious** and offbeat humour abounds, **and it creates a deep** connection with the audience.
(Label: Positive)

- By minimally editing an example, we control for pretty much all of the possible shortcuts that apply to the original.
- E.g., [summary starts with "*Hardly*" -> negative] is a pattern that could not hold anymore

Gardner et al. (2020)

Contrast Sets

Dataset	# Examples	# Sets	Model	Original Test	Contrast		Consistency
NLVR2	994	479	LXMERT	76.4	61.1	(–15.3)	30.1
IMDb	488	488	BERT	93.8	84.2	(–9.6)	77.8
MATRES	401	239	CogCompTime2.0	73.2	63.3	(–9.9)	40.6
UD English	150	150	Biaffine + ELMo	64.7	46.0	(–18.7)	17.3
PERSPECTRUM	217	217	RoBERTa	90.3	85.7	(–4.6)	78.8
DROP	947	623	MTMSN	79.9	54.2	(–25.7)	39.0
QUOREF	700	415	XLNet-QA	70.5	55.4	(–15.1)	29.9
ROPES	974	974	RoBERTa	47.7	32.5	(–15.2)	17.6
BoolQ	339	70	RoBERTa	86.1	71.1	(–15.0)	59.0
MC-TACO	646	646	RoBERTa	38.0	14.0	(–24.0)	8.0

Gardner et al. (2020)

Solutions

Broad Solutions

- Most solutions involve changing what data is trained on
  - Hard subset
  - Soft subset
  - Superset: add adversarially-constructed data, contrast sets, etc.
- For subsets: what do we train on?
  - Don't train on stuff that allows you to cheat
  - Train on examples that teach the real task rather than shortcuts

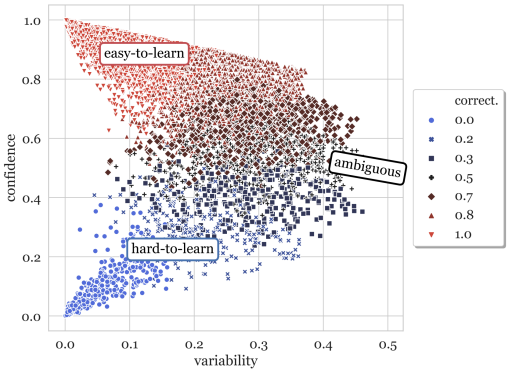
Dataset Cartography

- What happens with each particular example during training?
- Spurious correlations are *easy to learn*: a model should learn these early and always get them right
- Imagine a very challenging example
  - Model prediction may change a lot as it learns this example, may be variable in its predictions
- Imagine a mislabeled example
  - Probably just always wrong unless it gets overfit

Swayamdipta et al. (2021)

Data Maps

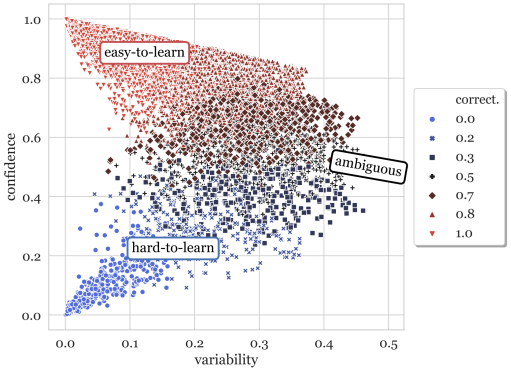
- Confidence: mean probability of correct label
- Variability: standard deviation in probability of the correct label
- Ambiguous examples: possible learnable (model knows it sometimes but not other times), but hard!



Swayamdipta et al. (2021)

Data Maps

- What to do with them?
- Training on hard-to-learn or ambiguous examples leads to better performance out-of-domain



Swayamdipta et al. (2021)

Debiasing

- Other ways to identify easy examples other than data maps
- Train some kind of a weak model and discount examples that it fits easily

one-hot label vector

log probability of each label

$$\mathcal{L}(\theta_d) = -(1 - p_b^{(i,c)}) y^{(i)} \cdot \log p_d$$

probability under a copy of the model trained for a few epochs on a small subset of data (bad model)

Utama et al. (2020)

Debiasing

Method	MNLI (Acc.) dev	*HANS*	Δ
BERT-base	84.5	61.5	-
Reweighting known-bias	83.5‡	69.2‡	+7.7
Reweighting self-debias	81.4	68.6	+7.1
Reweighting ♠ self-debias	82.3	69.7	**+8.2**

- On the challenging HANS test set for NLI, this debiasing improves performance substantially
- In-domain MNLI performance goes down

Utama et al. (2020)

Debiasing

- Other work has explored similar approaches using a known bias model

$$\hat{p_i} = softmax(\log(p_i) + \log(b_i))$$

probabilities from learned bias model — like the weak model from Utama et al. (prev. slides), but you define its structure

- *Ensembles* the weak model with the model you actually learn.
- Your actual model learns the *residuals* of the weak model: the difference between the weak model's output distribution and the target distribution.
- This lets it avoid learning the weak model's biases!

He et al. (2019), Clark et al. (2019)

Core Principles

- By reweighting data or changing the training paradigm, you can learn a model that generalizes better
- Most gains will show up **out-of-domain**. Very hard to get substantial improvements on the same dataset, unless you consider small subsets of examples (e.g., the toughest 1% of examples by some measure)

Final Project
(see spec and GitHub)