ACL 2020 Highlights

July 22, 2020

What's new

- Two new tracks:
  - Ethics and NLP
  - Interpretability and Analysis of Models for NLP
- Theme - Taking Stock of Where We've Been and Where We're Going
  - Reflect on field's progress and provide holistic view of current (w.r.t. past)
  - Identify limitations of SOTA models and impactful ideas to improve models
  - Bring novel ideas for advancing the field, e.g., to enable and measure a machine's ability in language processing beyond laboratory benchmarks
- Fully online
  - Pros: Detailed discussion, Coverage, Accessible content, Registration
  - Cons: Random meetings*

https://medium.com/@yoav.goldberg/the-missing-pieces-in-virtual-acl-a05327cf9a18

Disclaimer

- Generation
- Interpretability and Analysis of Models for NLP
- CSS and Social Media
- Semantics
- Resources and Evaluation
- Theme

Tutorial: Interpretability and Analysis in Neural NLP

- Better understanding
  - Better systems
  - More accountable systems
  - More interpretable models
- Techniques
  - Structural analyzes
  - Behavioral analyses
  - Interaction + Visualization (skipped)

Analysis Questionnaire

What is the goal of the study?

Pedagogical / Debugging / Debiasing / …

Understanding model structure / model decisions / data / ...

How do you quantify an outcome?

Who is your user or target group?

ML or NLP Expert/ Domain Expert / Student / Lay User of the System ...

How much domain/ model knowledge do they have?

Structural analyzes

- Questions:
  - What is the role of different components?
  - What kind of information do different components capture?
  - Does component A know something about property B?
- E.g.: Identifying layers in BERT that capture PoS

Structural analyzes

What is the goal of the study?

Scientific / Pedagogical / **Debugging** / Debiasing / **…**

Understanding model structure / model decisions / data / ...

How do you quantify an outcome? **Performance comparisons**

Who is your user or target group?

ML or NLP Expert / Domain Expert / Student / Lay User of the System ...

How much domain/ model knowledge do they have? **Enough to understand the model and problem domain**

Limitations:

- Probe complexity vs. Probe quality
- Correlation vs. Causation

Behavioral analyzes

- Challenge sets to target specific phenomena
  - Make inferences about model's representation based on model's behavior
- E.g.: Negation test
- Limitations
  - Limited coverage of tasks and languages
  - Hard to design
  - Challenge sets can have artifacts
  - Risk of overfitting challenge sets
  - Few insights on why the model failed to solve the task

Behavioral analyzes

What is the goal of the tool?

Scientific / Pedagogical / **Debugging** / Debiasing / **…**

Understanding model structure / **model decisions** / data / ...

How do you quantify an outcome? **(Relative) accuracy across different challenge sets**

Who is your user?

ML or NLP Expert/ Domain Expert / Student / ...

How much domain/model knowledge do they have? **Knowledge of target phenomena, but no model knowledge**

The answers will inform the following implementation questions:

Does the tool require interaction with the model? With the data? **Model treated as a "black box"**

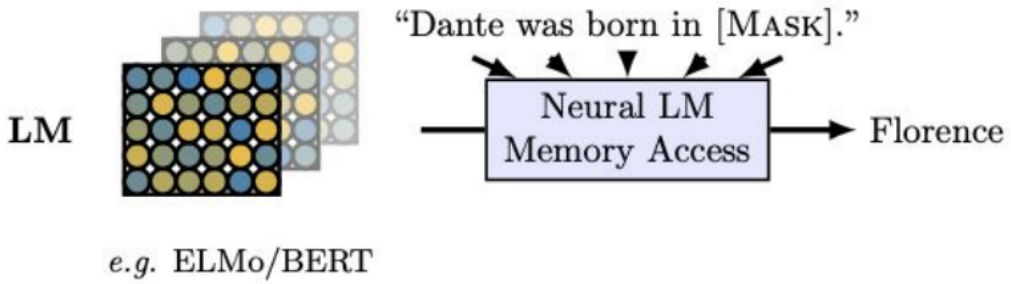
Can you change the model structure or model decisions? **No**

Open questions

- How can we make insights from these techniques actionable?
- What is the connection between representations' structure (measured by probing techniques) and model decisions (measured by challenge sets)?
- Can techniques like probing classifiers be adapted to measure something less correlational, and more causal?
- Want more? See EMNLP tutorial on Interpreting Predictions of NLP Models

Tutorial: Commonsense Reasoning for NLP

- Common Sense
  - Practical knowledge, reasoning concerning everyday situations/events
  - E.g.
    - ok to keep closet door open
    - not ok to keep fridge door open as food might go bad
- Essential for AI to understand human needs and actions better
- Do pre-trained LM already capture commonsense knowledge?



  - BERT performs well but all models perform poorly on many-to-many relations

Tutorial: Commonsense Reasoning for NLP

- Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?

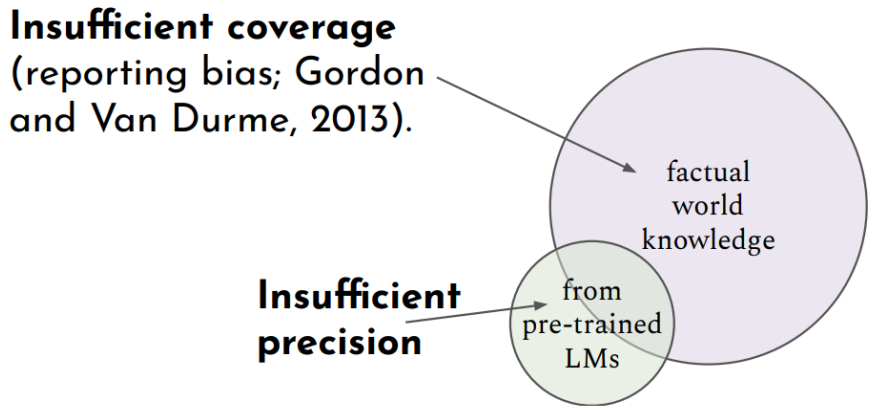
> A ____ has fur.
> A ____ has fur, is big, and has claws.
> A ____ has fur, is big, and has claws, has teeth, is an animal, …

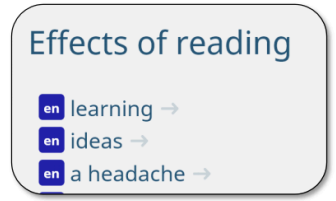
  - Perceptual (e.g. visual) < non-perceptual (e.g. encyclopaedic or functional) - can't be learned from texts alone
- Can you teach LMs symbolic reasoning? Reporting bias issue

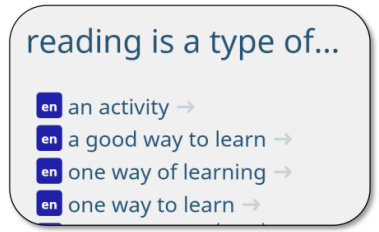
Always-Never: A chicken [MASK] has horns. A. never B. rarely C. sometimes D. often E. always

Tutorial: Commonsense Reasoning for NLP

- Pre-trained language models some commonsense knowledge - but it is far from an exhaustive source.
- Use with caution! LMs also generate false facts.
- Commonsense resources
  - ConceptNet – Semantic knowledge in natural language form

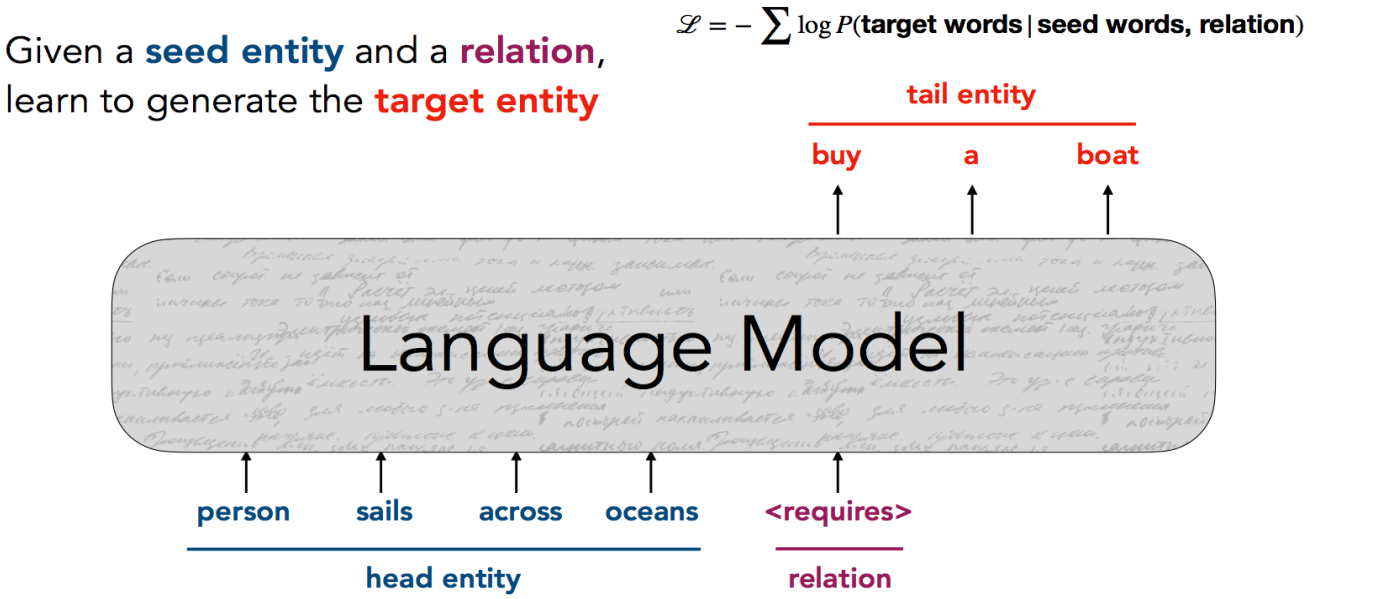






Tutorial: Commonsense Reasoning for NLP

- How to construct KBs automatically?

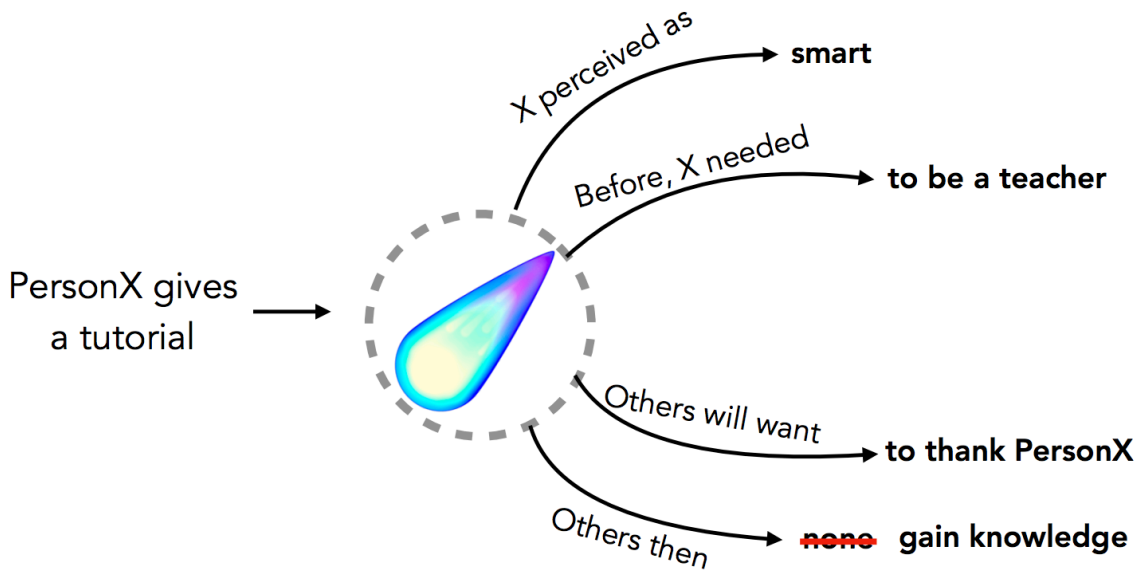


(Bosselut et al., 2019)

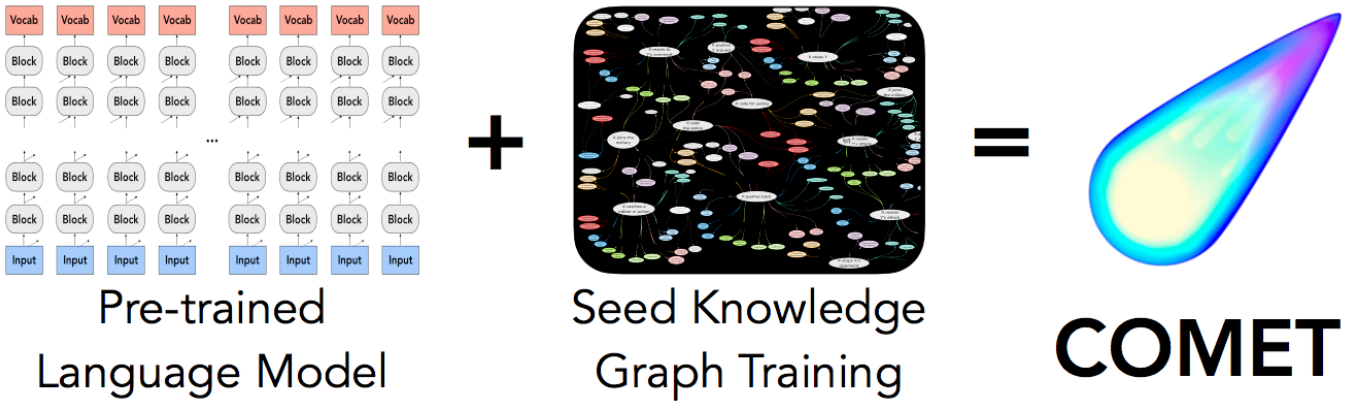
- Once fine-tuned, generate common sense knowledge for any input concept

Tutorial: Commonsense Reasoning for NLP

- How to construct KBs automatically?



- Converting LM to KM



Rewriting the Past: Assessing the Field through the Lens of Language Generation - Prof. Kathleen R. McKeown / Columbia University, New York

- NLG model can lie, might not plan longer text, might have no purpose
- General purpose model -> a model that works well for 1 task
- Address tasks that really matter (e.g. single document news summarization)
- Learn the task and not the dataset
- Bring language back to NLP
  - Analyze your output
  - Careful preparation and analysis of data sets

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

- Problem: Held-out accuracy can overestimate model performance
- Solution: Treat NLP model like a software, create test cases and prepare bug report
- CHECKLIST
  - What to test: capabilities (e.g. Vocab + POS)
  - How to test:
    - MFT – Check a behavior within capability (e.g., sentences with neutral adjective)
    - INV – Label-preserving perturbations (e.g., replace neutral words with other)
    - DIR – Label-changing perturbation (e.g., add negative phrases, fails if sent change > 0.1)
  - Tooling: BERT fill-ins, visualizations, lexicons

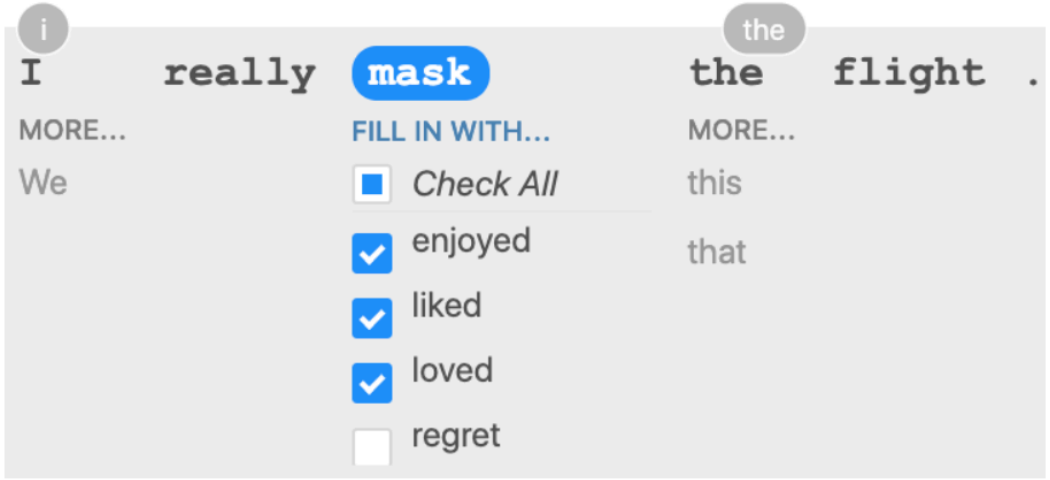
Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

	Test *TYPE* and Description	Failure Rate (%)					Example test cases & expected behavior
						RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	INV: Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not poor. pos or neutral It isn't a lousy customer service. pos or neutral
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is not private. neutral This is not an international flight. neutral
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plane would be awful, but it wasn't. pos or neutral I thought I would dislike that plane, but I didn't. pos or neutral
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, given it's a Tuesday, that this pilot was great. neg I don't think, given my history with airplanes, that this is an amazing staff. neg

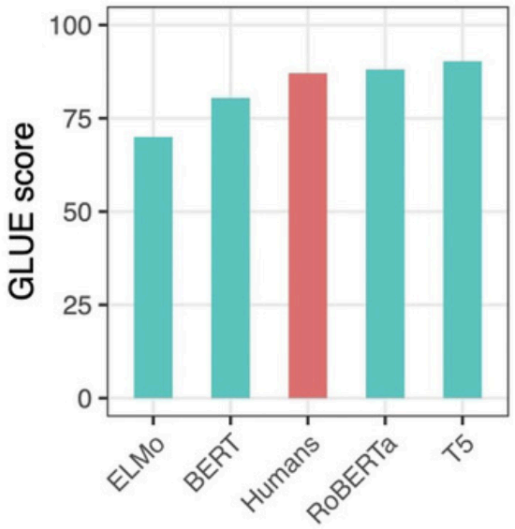
Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

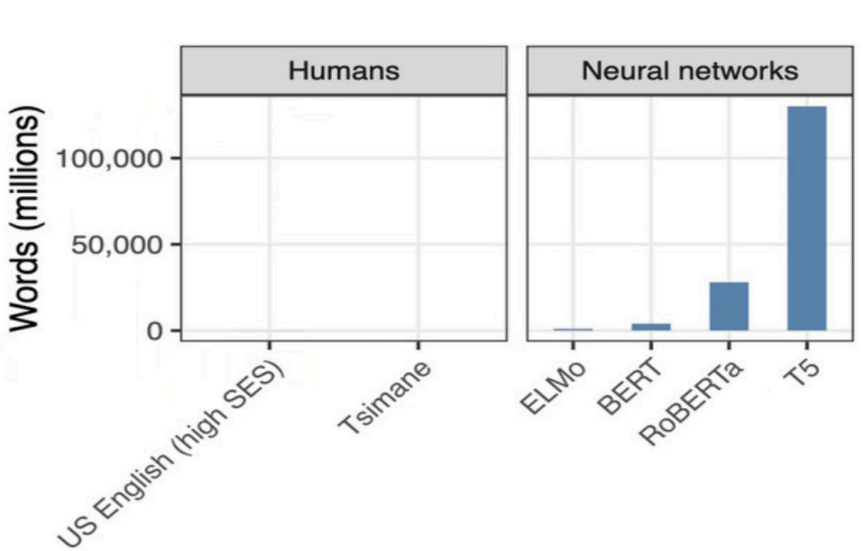
- Generating Test Cases at Scale
  - Templates: Relies on human creativity
    - “I didn’t love the food.” -> “I {NEGATION} {POS_VERB} the {THING}.”
    - {NEGATION} = {didn’t, can’t say I, ...}
    - {POS_VERB} = {love, like, ...}
  - Expanding Templates: Ask RoBERTa to fill-in

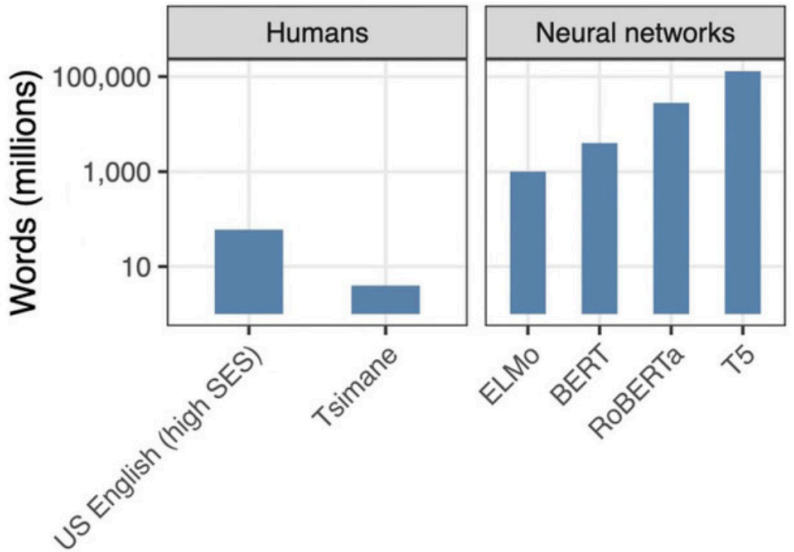


How Can We Accelerate Progress Towards Human-like Linguistic Generalization?

- Problem: BERT like models are (pretraining) sample inefficient and do not generalize like humans
- Solution: Evaluate for human-like generalization
- BERT like models do well,
  - But are pretraining-data hungry







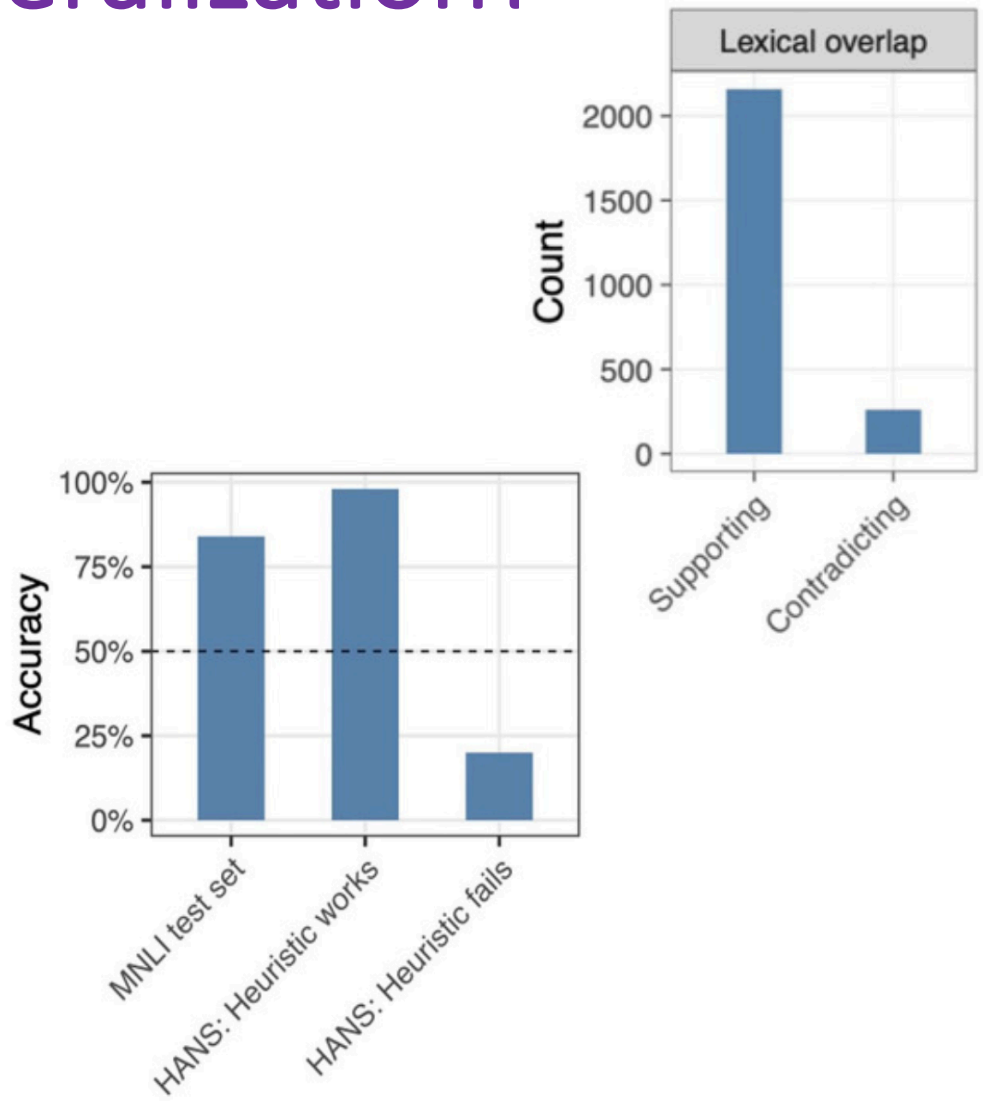
How Can We Accelerate Progress Towards Human-like Linguistic Generalization?

- Example application: NLI

The tall man is sleeping on the couch. →
The man is on the couch.

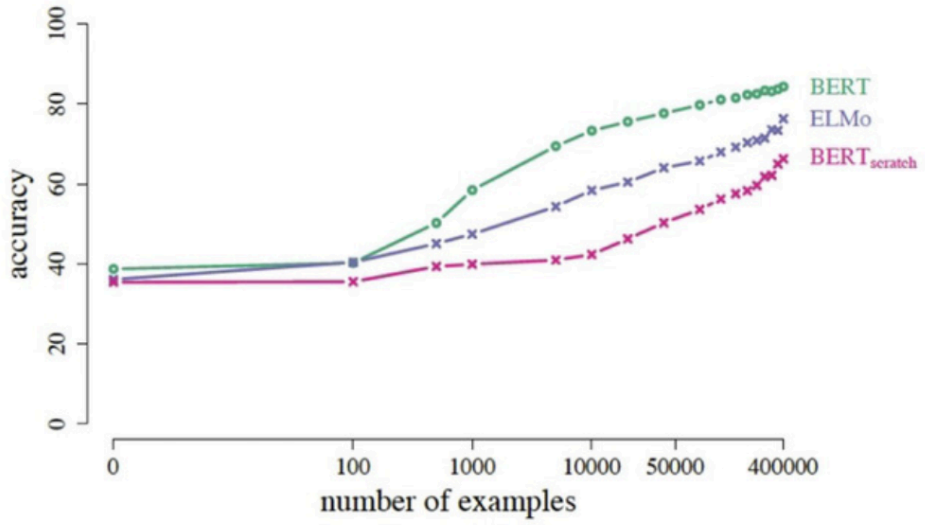
1. The judge was paid by the lawyer → The lawyer paid the judge.

2. The doctor was paid by the actor. ↛ The doctor paid the actor.



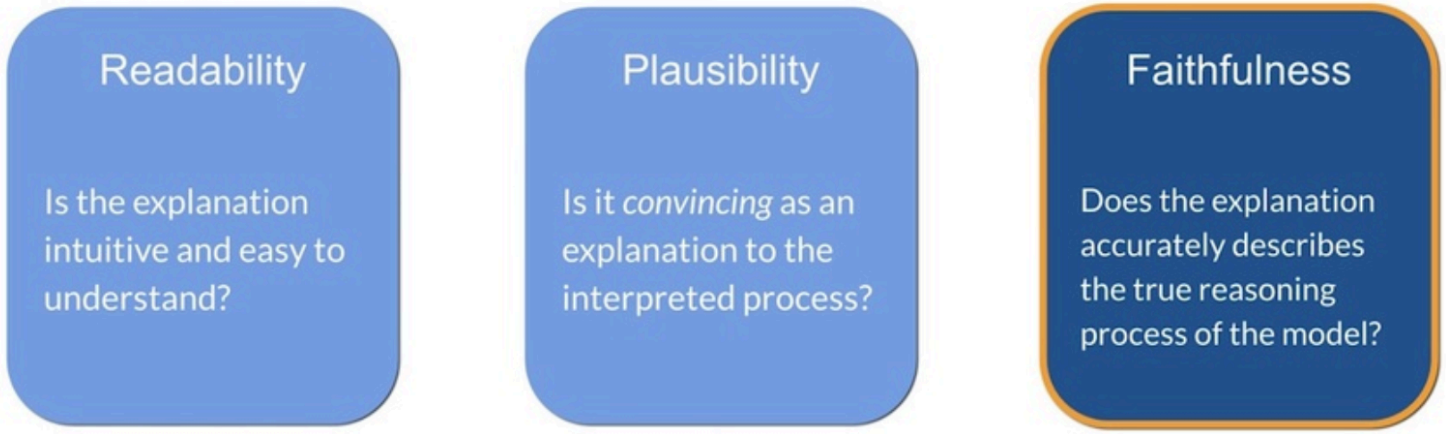
How Can We Accelerate Progress Towards Human-like Linguistic Generalization?

- How can we measure progress towards robust generalization from less data?
  - Standardized, moderately sized pretraining corpora
  - Expert-created evaluation sets that are inaccessible during fine-tuning
  - Few-shot learning



Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

- Problem: What is faithful explanation of a model's decision?
- Issue: Lot of research, but the foundations are still shaky
- What makes an interpretation useful?



- Tradeoff between them: Raw activations of NN are faithful, but not-readable.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

- How not to evaluate faithfulness
  - Faithfulness ≠ Plausibility
    - Many work conflate evaluating faithfulness and evaluating plausibility.
    - Plausible but unfaithful interpretation is akin to lying
  - Model decision process ≠ human decision process
    - Many work evaluate faithfulness by asking humans to rate explanation quality.
    - We (humans) can't understand models that need interpretation (otherwise, why research this?)
    - Evaluating interpretations using humans input is evaluating plausibility.
  - Don't trust untested claims of "inherent interpretability" of models

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

- How to evaluate faithfulness
  - Be explicit in what you evaluate
  - Faithfulness evaluation should not involve human-judgement on the quality of interpretation
  - Faithfulness evaluation should not involve human-provided gold labels.
  - Do not trust “inherent interpretability” claims.
  - Faithfulness evaluation of IUI systems should not rely on user performance.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

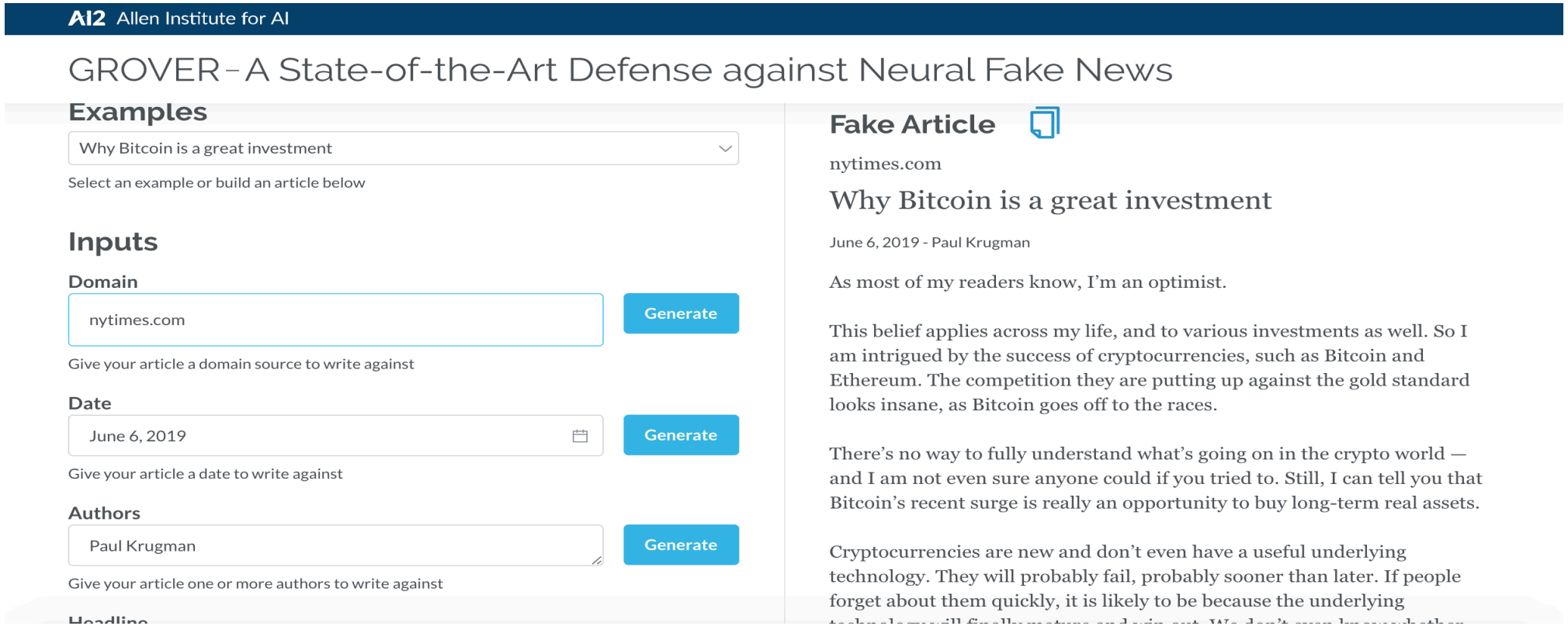
- How existing work has defined faithfulness
  - **The Model Assumption** - Two models will make the same predictions if and only if they use the same reasoning process.
    - An interpretation system is unfaithful if it results in different interpretations of models that make the same decisions.
  - **The Prediction Assumption** - On similar inputs, the model makes similar decisions if and only if its reasoning is similar.
    - An interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs.
  - **The Linearity Assumption** - Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.
    - Under certain circumstances, heatmap interpretations can be faithful.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

- Is faithful interpretation (binary) impossible?
  - Clear trend of proof via counter-example, for various interpretation methods, that they are not globally faithful.
  - Unproductive claim, trend is natural as an interpretation functions as an approximation of the model or decision's true reasoning process, so it loses information
- Possible direction – Evaluate faithfulness in "grayscale"
  - Domain restrictions - We should care about how the interpretation method behaves on natural input spaces and specific tasks.
  - Targeted interpretations - Can we define testable conditions on input examples that guarantee faithfulness of a method on them?

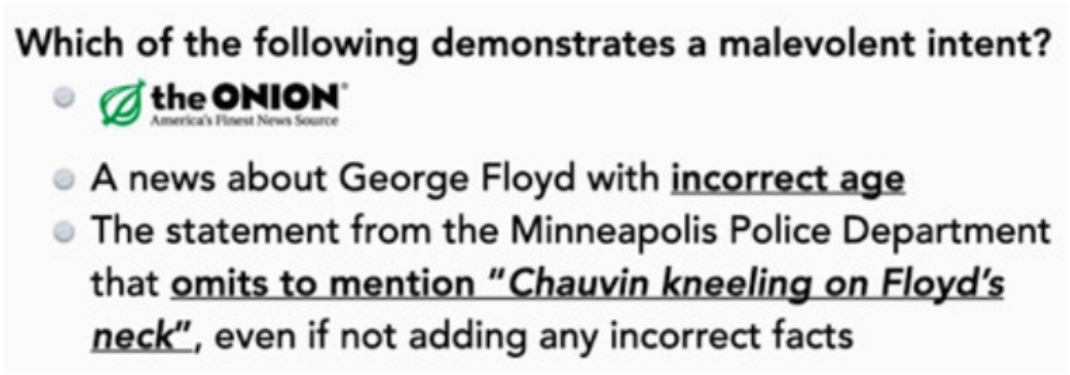
Fake Fake News and Real Fake News by Yejin Choi – FEVER workshop

- Are neural fake news as good as (or as bad as) REAL fake news?
  - GROVER can write fake news that are trustworthy to humans than human written fake news



Fake Fake News and Real Fake News by Yejin Choi – FEVER workshop

- Are all fake news bad and all REAL news good?
  - It is the underlying intent and the purpose that matter



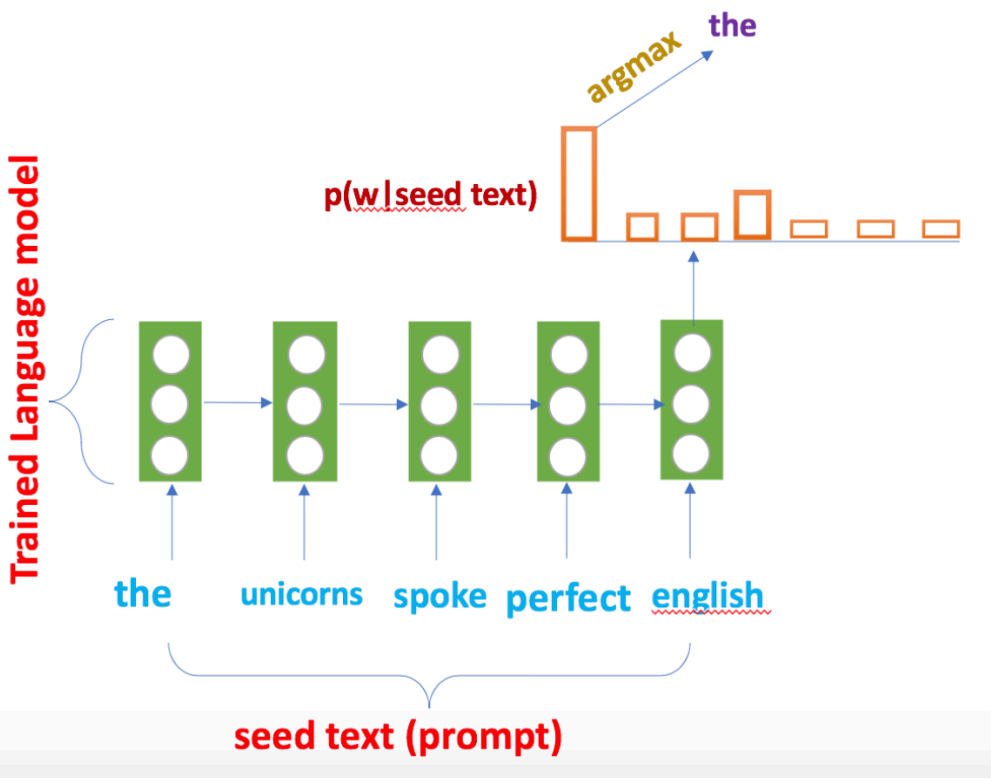
  - It is WHY not just WHAT that matters
- Three layers of fake news / manipulated media detection:
  - Perceptual/distributional level — **stylometric fingerprints**
  - Semantic level — **fact-checking**
  - Pragmatic level — **unfair or malevolent intent**

Fake Fake News and Real Fake News by Yejin Choi – FEVER workshop

- Good news – Neural fakes will have hard time to get the semantics and pragmatics right in mass-producing fake news
- Bad news – Neural model's aren't reliable for semantic and pragmatic analysis to stop misinformation
- All three levels needs to be focused.
- AI can't battle alone.
  - We need platform and system-based solutions as well.

Automatic Detection of Generated Text is Easiest when Humans are Fooled

- Problem: Detecting machine generated text from human written text
- Finding: Choice of decoding strategy impacts what the discriminator learns



w	p(w|context)
the	0.6
scientist	0.06
named	0.02
population	0.03
a	0.2
french	0.05
researcher	0.04

sort based on p

w	p(w|context)
the	0.6
a	0.2
scientist	0.06
french	0.05
researcher	0.04
population	0.03
named	0.02

Randomly sample a word from top-k words '[the, a]' based on probabilities 'softmax([0.6, 0.2])'

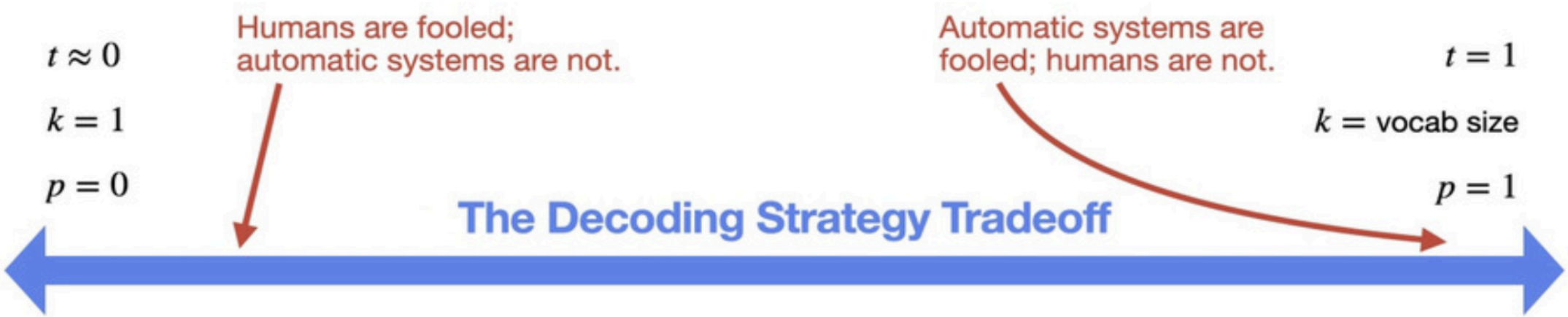
w	p(w|context)
the	0.6
scientist	0.06
named	0.02
population	0.03
a	0.2
french	0.05
researcher	0.04

sort based on p

w	p(w|context)	cum. p.
the	0.6	0.6
a	0.2	0.8
scientist	0.06	0.86
french	0.05	0.91
researcher	0.04	0.95
population	0.03	0.98
named	0.02	1.0

Randomly sample a word from top-p words '[the, a, scientist, french]' based on probabilities 'softmax([0.6, 0.2, 0.06, 0.05])'

Automatic Detection of Generated Text is Easiest when Humans are Fooled

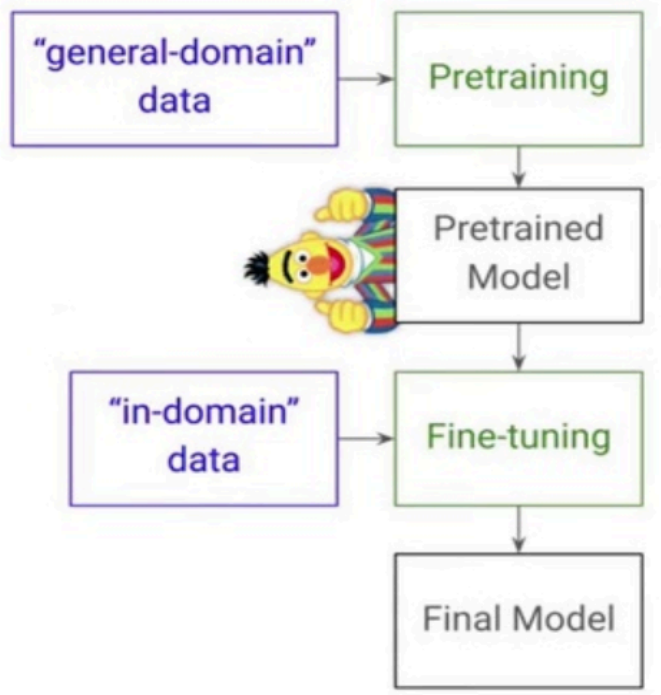


- Unlikely to ever choose uncommon words
- Text is not very diverse
- Text has an over-representation of common words.
- Generated text makes fewer semantic errors.

- Quite likely to choose uncommon words
- Text is very diverse
- Text looks distributionally similar to human-written text.
- Generated text has semantic errors that humans easily pick up on.

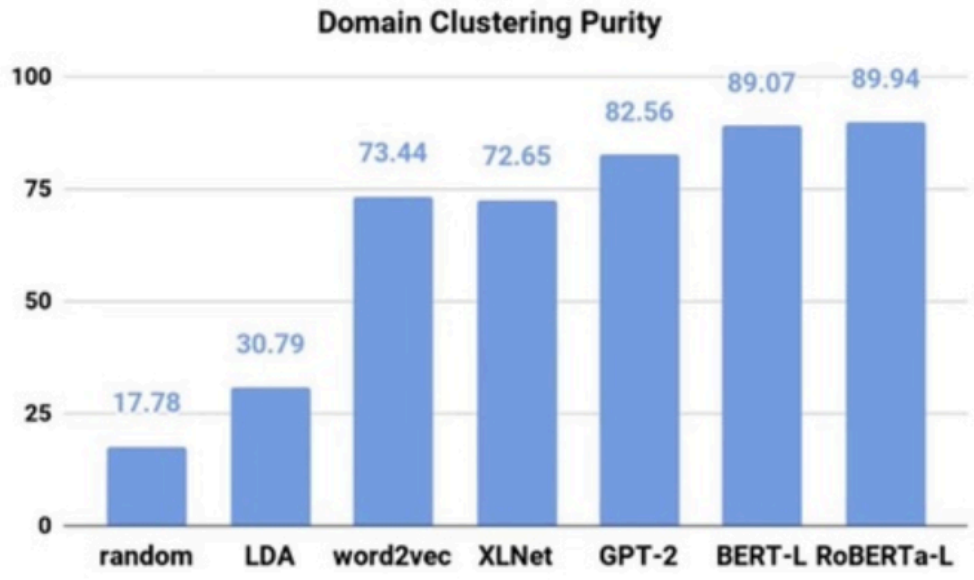
Unsupervised Domain Clusters in Pretrained Language Models

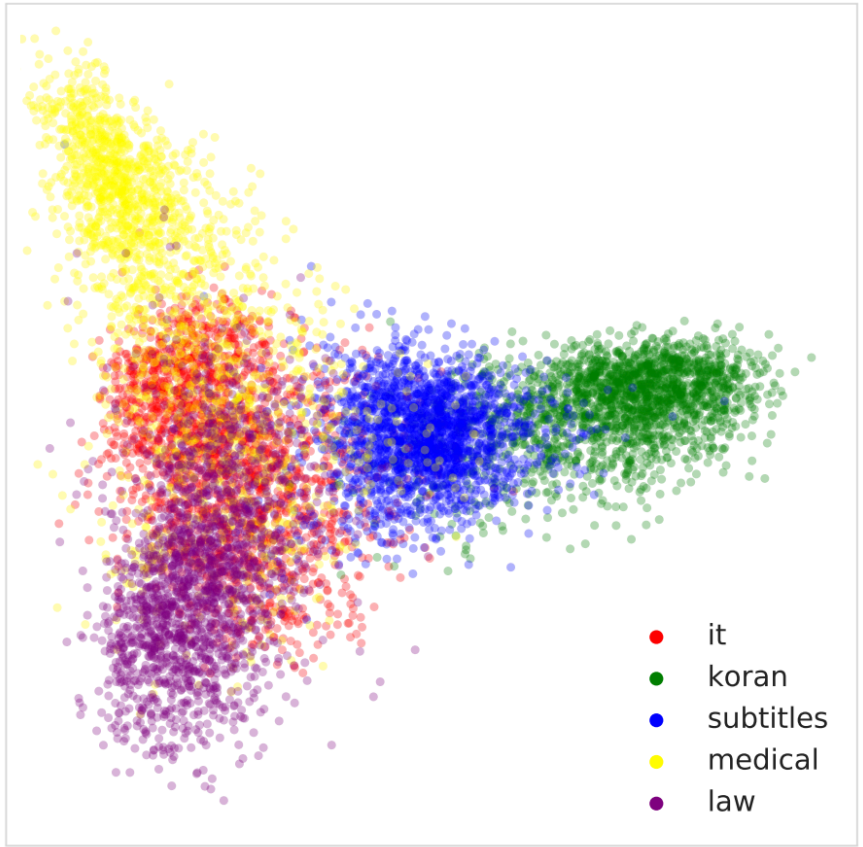
- Problem: NMT standard recipe, General-domain + In-domain training, but how do we define a domain?
- Solution: Data-driven approach to define domain, select in-domain data
- What's in a domain?
  - Topic? (sports/finance)
  - Data source? (TED, books)
  - Genre/style? (spoken, scientific)



Unsupervised Domain Clusters in Pretrained Language Models

- Hypothesis: LM induce implicit, “data-driven” domain clusters without domain supervision.
- LMs strongly encode domain related info





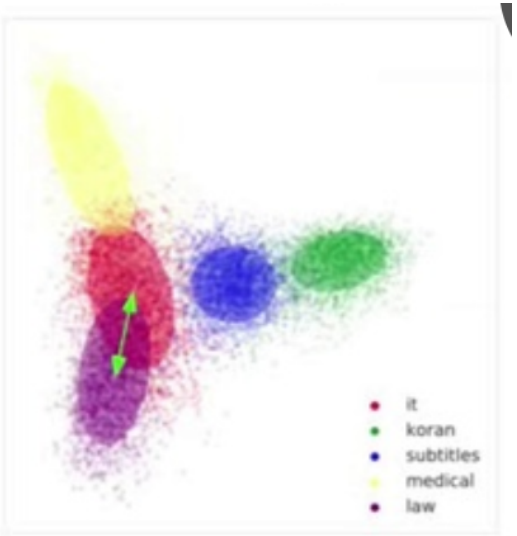
Unsupervised Domain Clusters in Pretrained Language Models

- Data-driven domain assignments can be better than naïve assignments based on the source

Law assigned to Medical	Law assigned to IT
- Viruses and virus-like organisms	”INFORMATION SOCIETY STATISTICS
where the glucose content is equal to or less than the fructose content.	This document must be attached to the certificate and field with it, except where there is a computerised checking system.

- Data selection for NMT

	Medical	Law	Koran	IT	Subtitles
Medical	**56.5**	18.3	1.9	11.4	4.3
Law	21.7	**59**	2.7	13.1	5.4
Koran	0.1	0.2	15.9	0.2	0.5
IT	14.9	9.6	2.8	**43**	8.6
Subtitles	7.9	5.5	6.4	8.5	27.3
All	53.3	57.2	**20.9**	42.1	**27.6**



Unsupervised Domain Clusters in Pretrained Language Models

- Data selection methods:
  - Domain-cosine: Compute centroid of in-domain data, select nearest-k examples around it
  - Domain-finetune: Fine-tune a pretrained LM for in-domain vs. out of domain (random) and select top-k according to probability or all positives
  - Domain-finetune + Pre-ranking: first run Domain-cosine, sample negative samples for Domain-finetune from farthest-k

	Medical	Law	Koran	IT	Subtitles	Average
Random-500k	49.8	53.3	18.5	37.5	25.5	36.92
Moore-Lewis-Top-500k	55	58	21.4	42.7	27.3	40.88
Domain-Cosine-Top-500k	52.7	58	**22**	42.5	27.1	40.46
Domain-Finetune-Top-500k	54.8	58.8	21.8	**43.5**	27.4	**41.26**
Domain-Finetune-Positive	55.3	58.7	19.2	42.5	27	40.54
Oracle	**56.5**	**59**	15.9	43	27.3	40.34
All	53.3	57.2	20.9	42.1	**27.6**	40.22

Other Highlights

- https://medium.com/analytics-vidhya/highlights-of-acl-2020-4ef9f27a4f0c
- https://medium.com/@lawrence.carolin/interpretability-and-analysis-of-models-for-nlp-e6b977ac1dc6
- https://towardsdatascience.com/knowledge-graphs-in-natural-language-processing-acl-2020-ebb1f0a6e0b1