

NAACL 2021 papers

🕒 Created At	@Jun 04, 2021 8:29 AM
🕒 Last Updated	@Jun 08, 2021 11:00 PM

Tutorials

Special Papers

Session 2D (New Challenges)

[Representing numbers in NLP: a survey and a vision](#)

[Implicitly abusive language — what does it actually look like and why are we not getting there?](#)

[The importance of modeling social factors of language](#)

[Preregistering NLP research](#)

[What will it take to fix benchmarking in NLU?](#)

Session 10E (New Challenges, etc.)

[Refining targeted syntactic evaluation of LMs](#)

[Adaptable and interpretable neural memory over symbolic knowledge](#)

Session 11E (Special Theme)

[A recipe for annotating grounded clarifications](#)

[Causal effects of linguistic properties](#)

[Translational NLP: A new paradigm and general principles for NLP research](#)

Papers in interpretability

Session 1B

[**Concealed data poisoning attacks on NLP models**](#)

[Mediators in determining what processing BERT performs first](#)

[**Automatic generation of contrast sets from scene graphs: probing the compositional consistency of GQA**](#)

[Do syntactic probes probe syntax? Experiments with Jabberwocky probing](#)

[Probing word translations in the Transformer and Trading Decoder for Encoder Layers](#)

Session 3C

[Generalization in instruction following systems](#)

[On attention redundancy: a comprehensive study](#)

[Towards interpreting and mitigating shortcut learning behavior of NLU models](#)

[Low-complexity probing via finding sub-networks](#)

[An empirical comparison of instance attribution methods for NLP](#)

[Does BERT pretrained on clinical notes reveal sensitive data?](#)

[Interpretability analysis for NER to understand system predictions and how they can improve](#)

Session 11B

[Topic model or topic twaddle? Re-evaluating semantic interpretability measures](#)

[Explaining neural network predictions on sentence pairs via learning word-group masks](#)

[Discourse probing of pretrained language models](#)

[Learning to learn to be right for the right reasons](#)

[Double perturbation: on the robustness of robustness and counterfactual bias evaluation](#)

[UniDrop: a simple yet effective technique to improve transformer without extra cost](#)

Session: Interpretability bird-of-feather social

Papers in linguistic theory, psycholinguistics

Session 12E

On biasing Transformer attention towards monotonicity

Finding concept-specific biases in form-meaning associations

Ab Antiquo: neuro proto-language reconstruction

How (non-)optimal is the lexicon?

Linguistic complexity loss in text-based therapy

Word complexity is in the eye of the beholder

Language in a (search) box: grounding language learning in real-world human-machine interaction

Papers in Computational Social Science

Session 6E

The structure of online social networks modulates the rate of lexical change

Framing unpacked: a semi-supervised interpretable multi-view model of media frames

Modeling framing in immigration discourse on social media

Automatic classification of neutralization techniques in the narrative of climate change scepticism

WikiTalkEdit: a dataset for modeling editors' behaviors on Wikipedia

Session 7A

What about the precedent: an information-theoretic analysis of common law

Characterizing English variation across social media communities with BERT

Session 7B (Green NLP)

It's not just size that matters: small LMs are also few-shot learners

Static embeddings as efficient knowledge bases?

Session 11A (ethics)

On the impact of random seeds on the fairness of clinical classifiers

Dynamically disentangling social bias from task-oriented representations with adversarial attack

An empirical investigation of bias in the multimodal analysis of financial earnings calls

Beyond fair pay: ethical implications of NLP crowdsourcing

Case study: deontological ethics in NLP

On transferability of bias mitigation effects in language model fine-tuning

Privacy regularization: joint privacy-utility optimization in LMs

Papers in semantics

Session 1E (sentence-level, textual inference)

Unifying cross-lingual SRL with heterogeneous linguistic resources

Meta-learning for domain generalization in semantic parsing

Session 4C (sentence-level, textual inference)

Understanding by understanding not: modeling negation in LMs

Disentangling semantics and syntax in sentence embeddings with pre-trained LMs

Temporal reasoning on implicit events from distant supervision

Session 5E (stylistic analysis)

Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa

Domain divergences: a survey and empirical analysis

Session 8C (sentence-level, textual inference)

Learning from executions for semantic parsing

[Compositional generalization for neural semantic parsing via span-level supervised attention](#)
[Incorporating external knowledge to enhance tabular reasoning](#)
[Game-theoretic vocab selection via the Shapley value and Banzhaf index](#)
[A flexible natural language interface for web navigation](#)

[Papers in discourse & pragmatics](#)

[Session 5B](#)

[Bridging anaphora resolution: making sense of the SOTA](#)
[Did they answer? Subjective acts and intents in conversational discourse](#)

[Session 12A](#)

[Predicting discourse trees from Transformer-based neural summarizers](#)
[Is incoherence surprising? Targeted evaluation of coherence prediction from LMs](#)
[Probing for bridging inference in Transformer LMs](#)
[Universal discourse representation structure parsing](#)
[Decontextualization: making sentences stand-alone](#)

[Papers in ML4NLP](#)

[Session 8A](#)

[Unified pre-training for program understanding and generation](#)
[How many data points is a prompt worth?](#)
[A primer in BERTology: What we know about how BERT works?](#)

[Session 9E](#)

[Grouping words with semantic diversity](#)
[Modeling content and context with deep relational learning](#)

[Session 14D](#)

[Revisiting simple neural probabilistic LMs](#)
[Limitations of autoregressive models and their alternatives](#)
[On the inductive bias of masked language modeling: from statistical to syntactic dependencies](#)

Tutorials

Interpretability tutorial: <https://github.com/hsajjad/Interpretability-Tutorial-NAACL2021>

Special Papers

Session 2D (New Challenges)

Representing numbers in NLP: a survey and a vision

Representing Numbers in NLP

Avijit Thawani and Jay Pujara and Pedro Szekely and Filip Ilievski

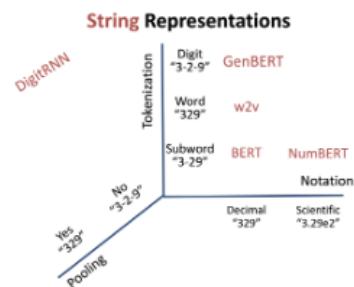
Information Sciences Institute

USCViterbi

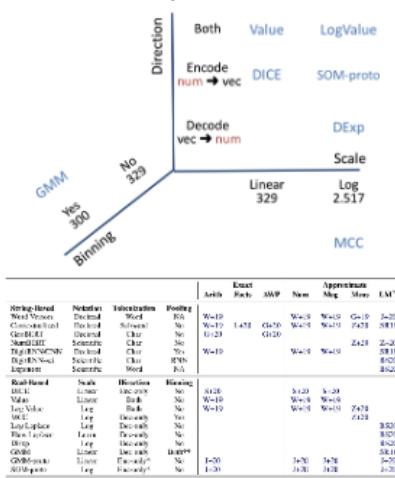
1. Taxonomy of Tasks

	Abstract	Grounded
Exact	$2 + 2 = 4$	2 balls + 2 balls = 4 balls Birds have two legs
Approx	'2' = 2.0 $4 > 2$	Tigers weigh 200 lbs

2. Taxonomy of Methods



Real Representations



3. Vision for Numeracy



1. Evaluation

	Abstract	Grounded
Exact	A%	B%
Approx	C%	D%

	Exact	Approx
$2 + 2 = 4$	<---->	200 lbs.
Bias	<---->	Variance
Real	<---->	String

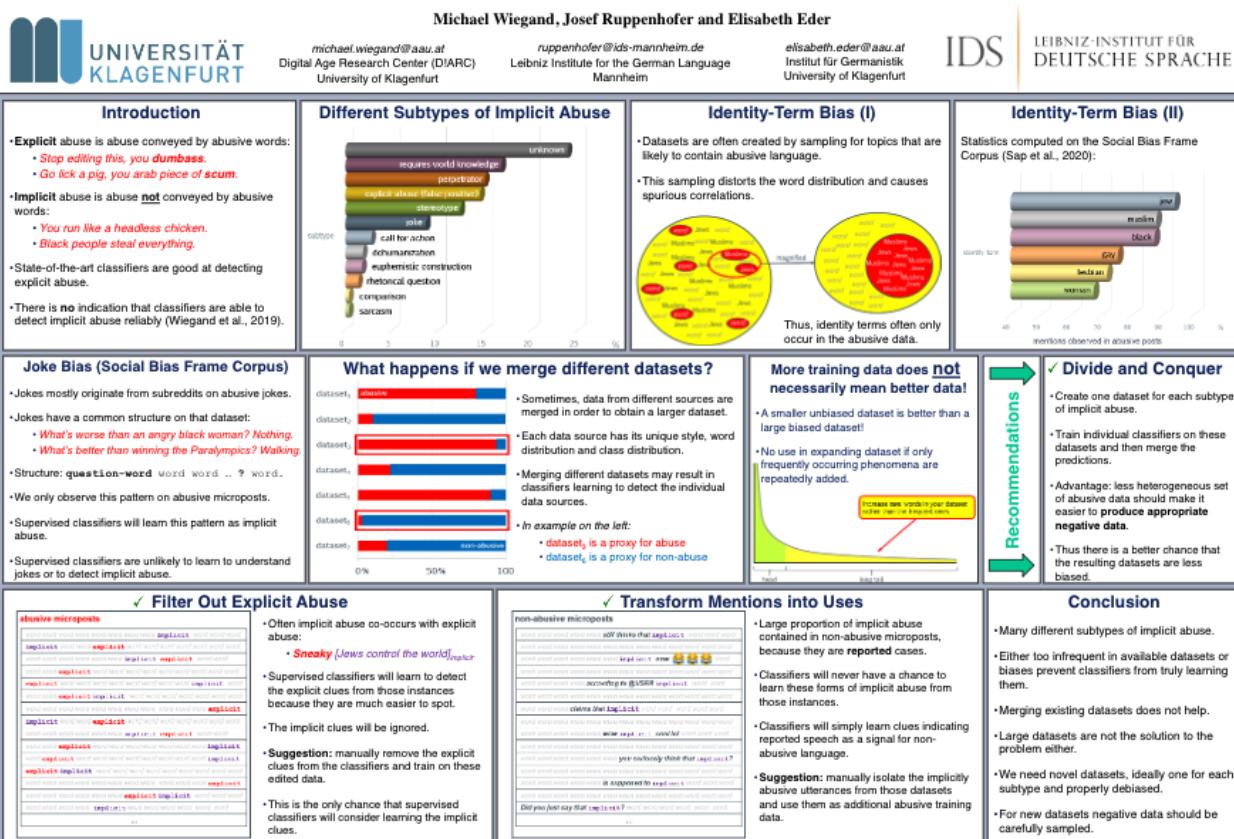
	Specificity	Expressivity
123.0	<---->	1.23, 1 ½, 123-125

2. Design Principles

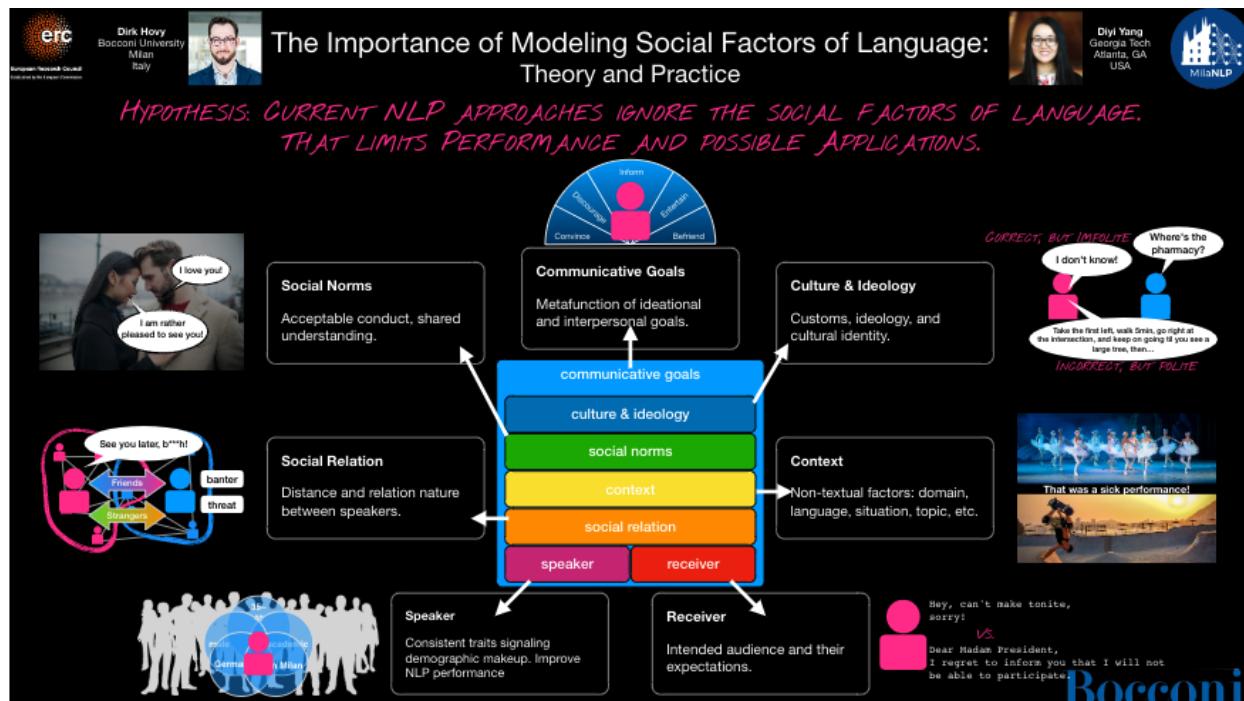


3. Broader Impacts

Implicitly Abusive Language – What does it actually look like and why are we not getting there?



The importance of modeling social factors of language



Preregistering NLP research

Preregistering NLP Research

Emiel van Miltenburg, Chris van der Lee, Emiel Krahmer

Preregistration is the practice of specifying what you are going to do, and what you expect to find in your study, before carrying out the study.

- Preregister a study by filling in a **preregistration form**.
- The form is **either public or private** (you decide).
- Preregistrations are **time-stamped** as evidence.

Why preregister?

- Reduce “researcher degrees of freedom.”
- Make your work more transparent:
 - What did you plan to do?
 - To what extent were you able to follow your plans?
 - Which findings are confirmatory/exploratory?

Registered reports are peer-reviewed preregistrations, that guarantee publication if the study has been carried out according to plan (and any changes are acknowledged).

- More constructive reviewing process.
- Less hassle to publish upon completion of the study.
- You can take your time! *Slow science* in NLP

We believe that **almost any NLP study could be preregistered**:

- ✓ Computationally-aided linguistic analysis
- ✓ NLP engineering experiment paper
- ✓ Reproduction
- ✓ Resource
- ✓ Survey Paper
- X Position

Open questions

- What should preregistration forms look like?
- Registered reports for *all* paper types?
- Could preregistrations form a separate publication type?
- ...

Contact

- @evanmiltenburg
- www.emielvanmiltenburg.nl
- c.w.j.vanmiltenburg@tilburguniversity.edu



What will it take to fix benchmarking in NLU?

- Position paper
- Goal: measure progress towards human-like language understanding in machines
- Problem: benchmarking for language understanding is broken
- Four criteria for building good benchmarks:
 - Validity: good performance on benchmark should imply robust in-domain performance on the task
 - Reliability: the labels in the test set should be correct and reproducible.
 - Statistical power: benchmarks should be able to detect qualitatively relevant performance differences between systems.
 - Social bias: benchmarks should reveal plausibly harmful social biases in systems, and shouldn't incentivize the creation of biased systems.

Session 10E (New Challenges, etc.)

Refining targeted syntactic evaluation of LMs



Refining Targeted Syntactic Evaluation of Language Models

Benjamin Newman, Kai-Siang Ang, Julia Gong, John Hewitt

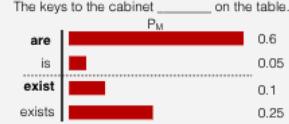


The Problem

- Understanding syntax underlies engineering and scientific applications of NLP systems
 - Engineering requires understanding models' **likely behavior** when sampling
 - Science requires models to have human-like **systematicity** of syntactic knowledge
 - Current evaluations (including **Targeted Syntactic Evaluation (TSE)**) (Marvin and Linzen, 2018) don't directly measure these
 - The use a small set of hand-selected verbs.
- Check models put higher probability on the grammatical of two sentences:
eg.^[1] The keys to the cabinet **are** on the table.
The keys to the cabinet **is** on the table.

Motivating Example

Consider:



are / is is the hand-selected verb in the minimal pair dataset

TSE

$P(\text{are}) > P(\text{is})$, so the score is 1.0

Likely Behavior

$P(\text{correctly conjugated verb}) = P(\text{are}) + P(\text{exist}) = 0.7$
TSE overestimates this because it assigns a binary score to each verb.

Systematicity

$P(\text{are}) > P(\text{is})$
 $P(\text{exist}) < P(\text{exists})$ ← because it's not in the minimal pair dataset.

Our Metrics

Consider:

- a minimal pair context c
- ℓ_+ is the correct conjugation and
- ℓ_- is the incorrect conjugation

Likely Behavior: Model Weighted Syntactic Evaluation (MW)

$$\text{MW} = \frac{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ | c)}{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ | c) + P_M(\ell_- | c)} \cdot \frac{P_M(\text{are}) + P_M(\text{exist})}{P_M(\text{are}) + P_M(\text{is}) + P_M(\text{exist}) + P_M(\text{exists})}$$

E.g.

$$= 0.7$$

Systematicity: Equally Weighted Syntactic Evaluation (EW)

$$\text{EW} = \sum_{\ell \in \mathcal{L}} \frac{1}{|\mathcal{L}|} [P_M(\ell_+ | c) > P_M(\ell_+ | c)] \cdot \frac{1}{2} (\mathbb{1}[P_M(\text{are}) > P_M(\text{is})] + \mathbb{1}[P_M(\text{exist}) > P_M(\text{exists})])$$

E.g.

$$= 0.5$$

Evaluations

Minimal Pairs Datasets

- Marvin and Linzen^[2]
- BLIMP^[3]

Lemmas: ~3500

- COCA^[4]
- Penn Treebank^[5]
- Giant Verb List^[6]

Models:

- BERT Large (cased)
- BERT Large (uncased)
- RoBERTa Large
- GPT2-XL

Templates	BERT cased			BERT uncased			RoBERTa			MW
	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE	
Simple	0.99	0.94	1.00	0.98	0.90	1.00	0.98	0.93	1.00	0.90
In a sentential complement	0.92	0.67	0.89	0.92	0.60	0.86	0.92	0.67	0.88	0.96
VP complement	0.91	0.68	0.90	0.90	0.65	0.90	0.91	0.68	0.90	0.89
Across subject relative clause	0.91	0.83	0.91	0.91	0.75	0.97	0.93	0.80	0.94	0.94
Across object relative clause	0.87	0.84	0.88	0.84	0.85	0.87	0.76	0.72	0.80	0.82
Across object relative clause (no that)	0.91	0.88	0.91	0.86	0.80	0.85	0.88	0.85	0.91	0.95
Across object relative (no that)	0.92	0.88	0.90	0.79	0.72	0.81	0.86	0.82	0.89	0.95
In object relative clause	0.93	0.95	0.97	0.91	0.97	0.99	0.89	0.91	0.97	0.91
In object relative (no that)	0.90	0.91	0.92	0.81	0.82	0.82	0.82	0.83	0.90	0.91
BLIMP	0.81	0.73	0.90	0.74	0.69	0.85	0.70	0.66	0.78	0.82

Qualitative Examples

The sentence that the sister are young	metric	encounters	oos	oos	encounters	we	glue	lacos	oos	enc
	0.26	0.027	0.027	0.048	0.021	0.020	0.018	0.012	0.012	0.016

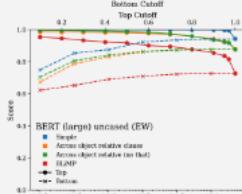
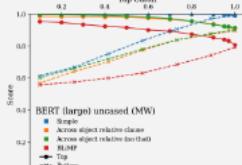
The plate that the executive are tall	metric	encounters	oos	oos	encounters	we	glue	lacos	oos	enc
	0.688	0.024	0.025	0.023	0.022	0.020	0.017	0.015	0.014	0.013

Likely Verbs

Why are our scores low?

We look at the top and bottom p% of models' distributions:

- Top: 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 97, 100%
- Bottom: 50, 10, 1, 0.1, 0.01, 0.001, 0.0001%



Conclusion

- We refine TSE to measure **likely behavior** (with MW) and **systematicity** (with EW) of language models
- We find models more often correctly conjugate verbs they deem likely

[1]Marvin, Linzen, Eisner, and Neel Guha. 2018. Assessing the utility of LSTMs to learn syntax-sensitive dependencies. ICLR.
[2]Marvin, Linzen, and Neel Guha. 2019. Targeted syntactic evaluation of language models. ANLP.
[3]Paul Vielliard, Alina Patashnik, Horia Bozdogan, Michael Lewis, David Pesetsky, and Ivan A. Rizzi. 2020. BLIMP.
[4]Paul Boag, 2006. The lexicon of contemporary American English.
[5]Mark Davies, 2008. The lexicon of contemporary American English.
[6]Peter Tiberius, 2004. The giant verb list. 2004. Building a large annotated corpus of English. The Penn Treebank, 4.
[7]Peter Tiberius, Wei Ding, Quirk & Stary, 2016. Stanford coref test: 3.2M coref pairs plus spelling rules and irregular verb lexicons.

Adaptable and interpretable neural memory over symbolic knowledge

Adaptable and Interpretable Neural Memory Over Symbolic Knowledge

Pat Verga*, Haitian Sun*, Livio Baldini Soares, William W. Cohen
 (patverga, liviobs, wcohen@google.com, haitians@cs.cmu.edu)

Google Research

Introduction

- Language models encode a large amount of factual knowledge but all of that information is stored in latent distributed representations
- Models become black boxes which lack interpretability - It's hard to know what the model knows
- Particularly important as models become bigger and bigger while being trained on larger chunks of the internet containing toxic information
- Cannot selectively remove or add information
- **Can we design models which are:**
 - 1- Interpretable by humans
 - 2- Accurate
 - 3- Updatable

Fact Injected Language Model (FILM)

- FILM extends an entity-augmented Transformer LM (Entities as Experts, aka EaE) that learns neural representation of entities, which are stored in an **Entity Memory**.

- FILM adds a **Fact Memory** - a key-value memory containing KG facts, constructed compositionally by combining neural entity representations from the **Entity Memory** with learned representations of relations. Memory access is pretrained from a (partially) entity-linked corpus (Wikipedia) with passages distantly-aligned with WikiData.

Results

Is our model accurate?

- FILM outperforms sota baselines on WebQuestionsSP - even T5-11B with 100x larger encoder and 13x more parameters **including FILM's memory**.
- Improvements are more dramatic for novel questions (no overlap) answerable from FILM's KG.
- FILM is also SOTA on other datasets (e.g., LAMA-TREx)

	Full Dataset: Total	Full Dataset: No Overlap	WikiData Answerable: Total	WikiData Answerable: No Overlap
FILM	54.7	36.4	78.1	72.2
EaE	47.4	25.1	62.4	42.9
T5-11B	49.7	31.8	61.0	48.5
BART Large	30.4	5.6	36.7	8.3
RAG	50.1	30.7	62.5	45.1
DPR	48.6	34.1	56.9	45.1

Can we inject new facts?

- FILM can utilize new facts injected at inference time without **any additional training**.

	Trained normally	Trained without passages relating question and answer entities	Trained on filtered passages + injected facts
FILM	56.5	38.7	48.0
EaE	45.8	28.8	-

Can we update stale memories?

- In a synthetic 'updated' world where answers to questions are replaced by type-consistent alternatives, FILM can utilize a set of corresponding updated facts to accurately answer those questions.

	FILM with old memory or updated facts	0.0
+ updated memory	54.5	

Conclusion

Interpretable knowledge is compatible with neural LMs

- FILM has 110M parameters for encoding and 720M parameters for memory
- On factual QA tasks, FILM dramatically outperforms much larger models
 - T5-11B 48.5% accuracy on answerable novel questions => FILM 72.2% accuracy
 - T5-11B has 100x the number of non-memory parameters
- The Fact Memory is compositionally defined and can be modified by **injecting new facts** or **editing old facts**

*equal contribution, work done at Google

Session 11E (Special Theme)

A recipe for annotating grounded clarifications

NAACL 2021 papers

8



A recipe for grounded clarifications



Luciana Benotti and Patrick Blackburn

Universidad Nacional de Córdoba, ARGENTINA and Roskilde University, DENMARK

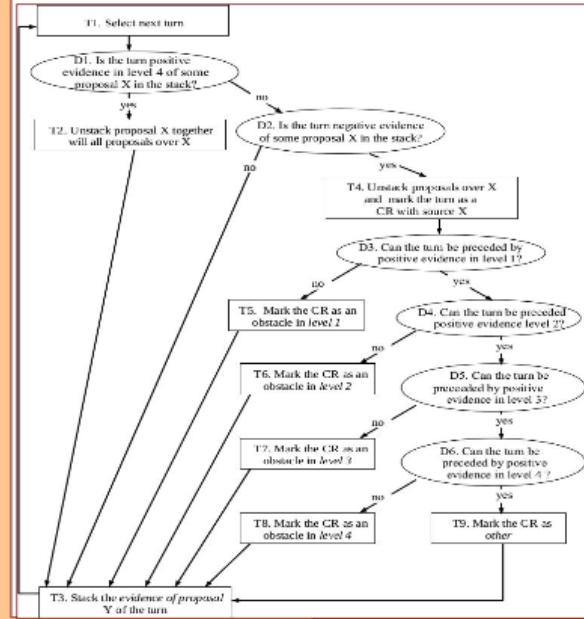
RESEARCH PROBLEM

1. To interpret the communicative intents of an utterance, we need to ground it in something outside language.



2. Clarification mechanisms make this interpretative process explicit - they ground utterances in world modalities such as vision, touch, movement,...
3. However form is not a robust indicator of clarification requests or clarification responses (Jurafsky 2006, Purver 2018) - we need something more sophisticated.
4. Here we propose a novel recipe for clarification annotations that unifies and extends previous accounts.

THE RECIPE FOR ANNOTATING GROUNDED CLARIFICATIONS



THE GROUNDED TEST

OK is ambiguous, it could mean:
Modality 4: OK, I did it (Object manipulation)

NGA in level 4: Do you want me to go above the carpenter?

Modality 3: OK, I see. (Vision)

NGA in level 3: I do not see the green bay

Modality 2: OK, I heard you. (Hearing)

NGA in level 2: The green what?

Modality 1: OK, so you want to talk to me. (Socioperception)

NGA in level 2: Are you talking to me?

Given an utterance U, a subsequent turn is a negative grounding act in modality m if it cannot be preceded by a positive grounding acts of U in m.



"Hi! We're reserving a table for Wednesday the 7th."



"For seven people?"

Causal effects of linguistic properties

Introduction

Our goal is to estimate the causal effects of linguistic properties on downstream behavioral outcomes.

For example,

- Does writing a complaint politely lead to a faster response time?
- How much will a positive product review increase sales?

Traditional neural networks and regressions rely on correlations to answer these questions. We want to make stronger causal conclusions and make three contributions towards this:

1. First, we formalize the causal quantity of interest and establish assumptions needed to identify this from observational data.
2. Second, we propose an estimator and prove bounds on its bias.
3. Last, we offer a concrete estimation algorithm.

Causal Inference Background

Causal inference from observational data is well-studied. In this setting, analysts are interested in the effect of a **treatment T** (e.g., a drug) on an **outcome Y** (e.g., disease progression) [1].

The average treatment effect (ATE) is a statistical estimand for measuring the causal effect of T on Y. It is,

$$\psi = E[Y; do(T=1)] - E[Y; do(T=0)]$$

where the operation $do(T=t)$ means that we hypothetically intervene and set the treatment T to some value.

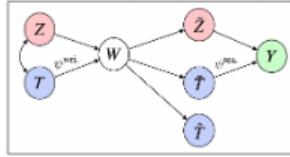
Typically, the ATE Ψ is not the simple difference in average conditional outcomes $E[Y|T=1] - E[Y|T=0]$. This is because confounding variables **C** are associated with both the treatment and outcome. When C is observed, we can compute the ATE as:

$$\psi = E_C[E[Y|T=1, C] - E[Y|T=0, C]]$$

i.e. group the data by C, calculating the average difference in outcomes between each group, then averaging over groups [2].

Formalizing the causal quantity of interest

We begin by proposing the following causal graphical model to describe the mechanism by which text influences outcomes.



A writer uses linguistic property T and other properties Z , which may be correlated (denoted by bi-directed arrow), to write the text W . From the text, the reader perceives the property of interest, captured by \hat{T} , and together with other perceived information Z , produces the outcome Y . In practice, one only has access to the variables W, Y , and a proxy for the treatment \hat{T} which corresponds to the predictions of a classifier or lexicon (e.g. a politeness or sentient classifier).

Our first result argues that the ATE obtained by imagining interventions on the readers perception \hat{T} is a good way to formalize the causal effects of textual properties.

Bounding the error

The section above is all well and good, but there's one big problem. The ATE we argue for is based on an unobserved variable: \hat{T} . Our second result says that the ATE obtained from intervening on the reader's perception \hat{T} is equal to the ATE obtained from intervening on the proxy label \hat{T} minus a term related to the error rate of the proxy. The error is a positive term, meaning our estimate only attenuates the true ATE:

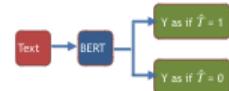
$$\psi^{\hat{T}} = \psi^T - error(\hat{T}, \hat{T})$$

This is a novel result for causal inference - prior work in the space required additional assumptions, namely access to an extra measurement model $P(\hat{T}|\hat{T})$ [3].

Estimation

Our last contribution is a concrete algorithm for estimating causal effects of linguistic properties. It has three stages:

1. Use a form of distant supervision to improve the quality of the proxy labels \hat{T} .
2. Train a neural network to predict both of Y's potential outcomes when \hat{T} is 0 and 1.

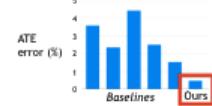


3. Run inference over a test set and calculate the average difference between the two potential outcomes.

Experiments

Our method gives high fidelity estimates and is practically useful

We experimented with two datasets: (1) a corpus of **Amazon reviews**, answering "what is the causal effect of sentiment on [simulated] sales?" and (2) real-world **financial complaints** [4], answering "what is the effect of complainant politeness on response time?"



References

1. Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Journal of the American Statistical Association* 78(359):496–512.
2. Judea Pearl. 2009. *Causality*. Cambridge University Press.
3. Zach Wood-Duchy, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of EMNLP*.
4. Naoki Egribo, Christian Fong, John Gummadi, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.05143*.

rpryzant@stanford.edu

Translational NLP: A new paradigm and general principles for NLP research

Translational NLP: A New Paradigm and General Principles for Natural Language Processing Research

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, Harry Hochheiser

NAACL 2021



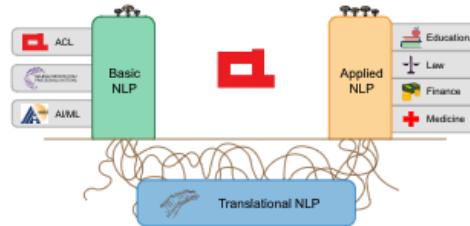
Carnegie
Mellon
University



National Institutes of Health
Funding support from
T15 LM005079

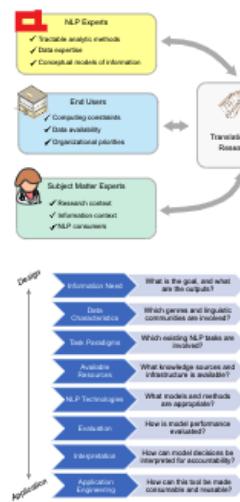
Abstract

- NLP research combines the study of universal principles (basic science) with applied science in specific use cases and settings.
- Translating basic innovations into successful applications, and finding research questions driven by applications, is not formally studied.
- Significant valuable work goes on underneath the surface, connecting basic and applied science through translational processes and translational questions.
- We present a general framework to help frame translational problems and design translational efforts, to improve successful exchange between basic and applied NLP.



Translational NLP is

- Application-driven solutions with generalizable impact
- Reusable processes and technologies to bridge between basic and applied science
- Already going on! But not formally studied.



Who's involved?

Translational NLP moves from systems to solutions: addressing information needs in real-world contexts.

Translational NLP Researchers bring together different stakeholders to design and manage NLP solutions.

What's involved?

We present eight questions to start the discussion around any translational NLP solution

Our questions are a starting point for evolving translational NLP development.

What does it look like?

Translational NLP projects have three components:

Problem

What's the need?



Solution

What was done?



Translational Impact

What did we learn for the next project?



For detailed examples, see our paper!

What's next?

Past

What translational research have we been doing and what have we learned?

Present

Learn from other places where NLP is being used and increase impact

Future

New, application-driven research questions
Building generalizable processes for translation

Papers in interpretability

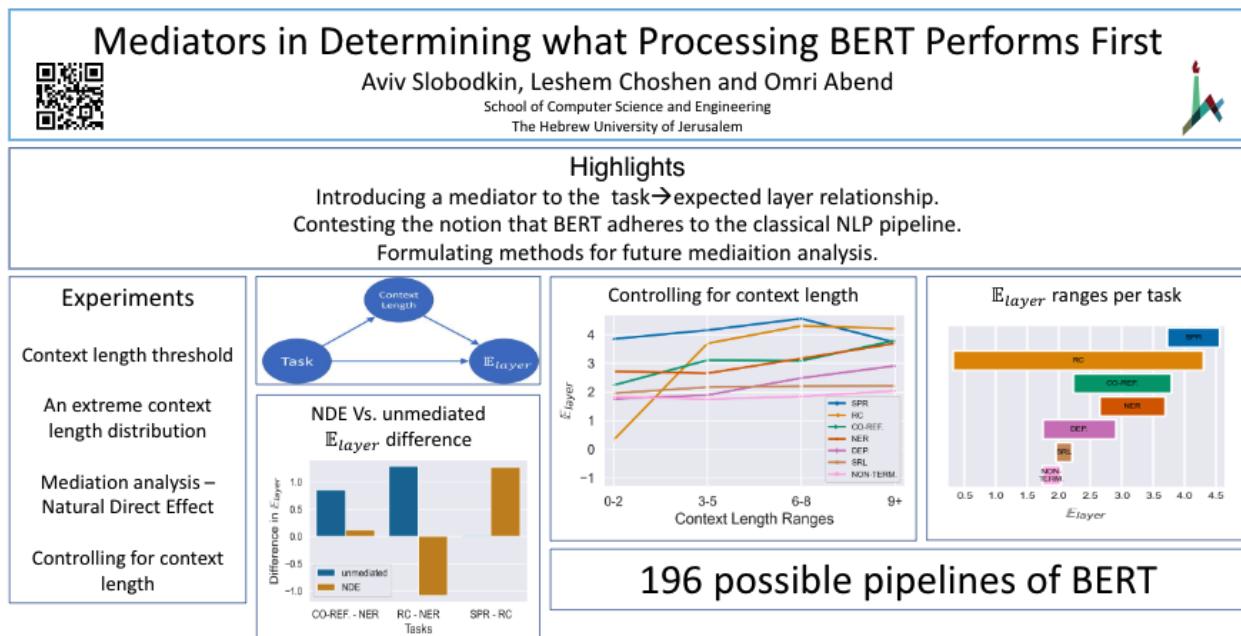
Session 1B

Concealed data poisoning attacks on NLP models

- Why scraping from internet can be a bad idea.
- Data poisoning attacks can turn any phrase (e.g., UC Berkeley) into a trigger for the negative class.
- This attack can be concealed.
- Crafting poison examples idea: use gradient of final prediction w.r.t poison example. Replace the e.g., "UC Berkeley" into something else.
- Too slow... approx: only do one step of training.
- Tasks:

- Sentiment: error rate on sentences with trigger phrase. Note: Regular validation accuracy is unaffected!
- Language models: measure how often LM generations are negative (with human evaluations) when generating "Apple iPhone". Finetune LM.
- Defending LM?
 - Defending with early stopping.
 - Identifying poison example using perplexity of a pretrained LM? This is hard.

Mediators in determining what processing BERT performs first



Automatic generation of contrast sets from scene graphs: probing the compositional consistency of GQA

- GQA dataset: for real-world graphs
 - Starting from (image, scene graph, Q, A), generate (image, scene graph, Q', A'), where Q' and A' are minimal changes from Q and A.
- Models struggle with our contrast set.
- Training on perturbed set leads to more robust models. Augment additional ~80k example (about 8%)

- Can measure the contrast consistency of the contrast set.

Do syntactic probes probe syntax? Experiments with Jabberwocky probing



Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing

Rowan Hall Maudslay Ryan Cotterell

[1/7] Syntactic Probing investigates whether unsupervised models implicitly learn syntax in their representations.

[2/7] It does this by training a supervised model to predict syntax using another model's hidden state; if it can do this, then people argue those representations contain syntax.

[3/7] The probing literature is frequently cited to support the claim that models like BERT do encode syntax.

[4/7] The trouble is, the sentences used for probing are real-world sentences, which do not properly isolate syntax. Chomsky:

[5/7] To investigate whether probes leverage semantic patterns to aide in their syntactic predictions, we create an evaluation corpus of syntactically parseable but semantically nonsense Jabberwocky sentences.

[6/7] For two probes, we find that performance in this setting drops (>50%), but in most cases remains above baselines.

[7/7] This begs the question: what scores constitute "knowing syntax"?

- RQ: Does this encode syntactic structures?
- Motivation: Chomsky argued that syntax and semantics are separate. What's the difference between a syntactic probe and a parser? (Hall Maudslay et al., 2020)
- Develop "jabberwocky sentences" test set. Substitute words into nonsense words.
 - On these sentences, the probe performance dropped, showing they use semantic confounds to predict syntax.
 - ... which raises the question: what scores would constitute "knowing syntax"?
- Finally: note that probes often adopt a simplified definition of syntax, vastly reducing search space.

Probing word translations in the Transformer and Trading Decoder for Encoder Layers

- Detect word translation in encoder and decoder layers.

- Show that word translation already happens in encoder layers.
 - By probing: encoder layers can give 40-50 acc. Decoder layers have between 16 to 67 acc.
- Given that Transformer encoder layers non-autoregressively perform word translation, we find that balancing between non-autoregressive translation and autoregressive translation can be achieved simply by adjusting encoder and decoder depth
 - Increasing encoder depth while decreasing decoder depth can increase decoding speed with improved translation quality.

Session 3C

Generalization in instruction following systems

<https://underline.io/events/122/sessions/4139/lecture/19865-generalization-in-instruction-following-systems>

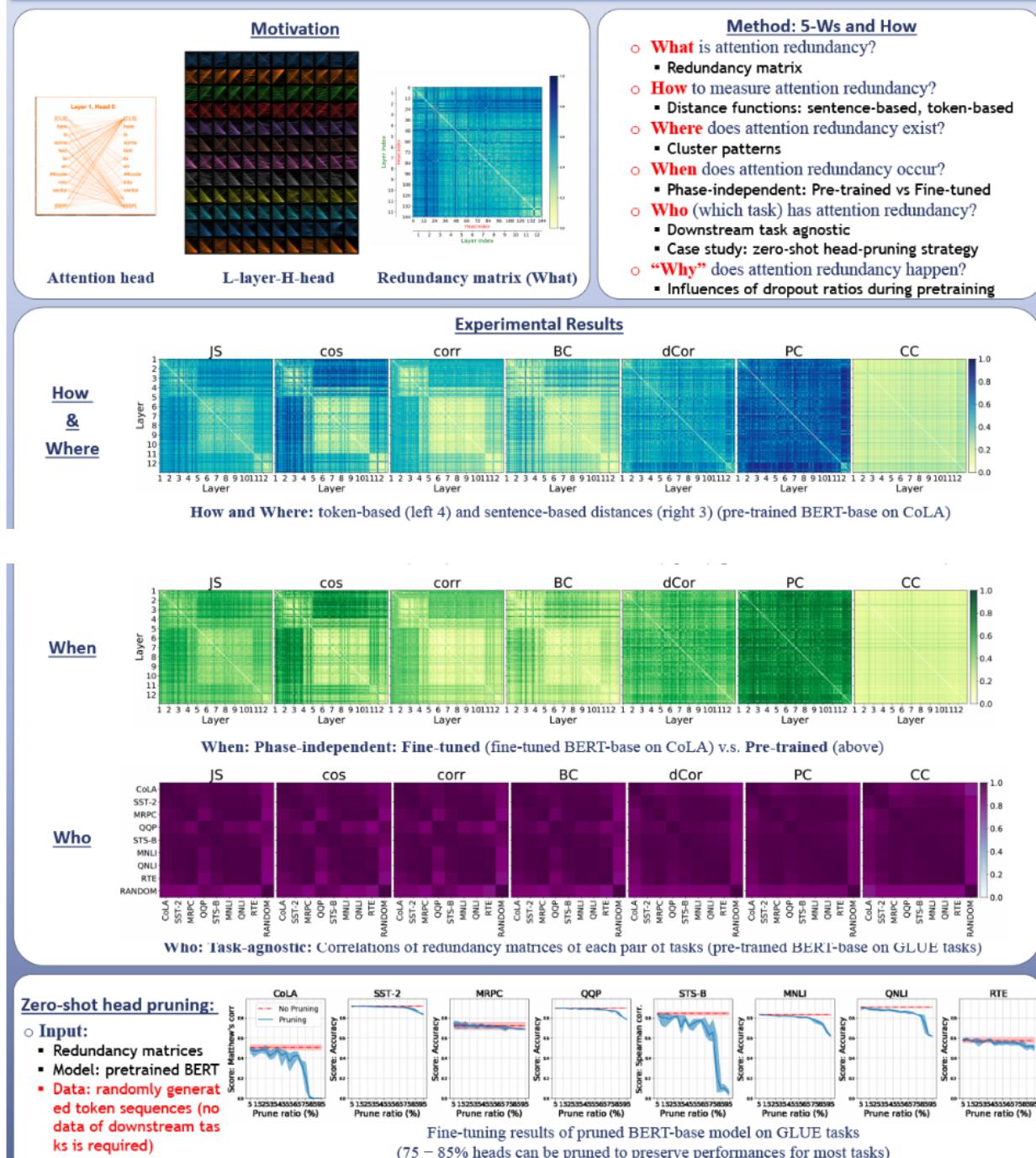
- Task: Given a configuration, move a block from the source location to the target location.
- Aim to understand if the test performance of these models indicates an understanding of the spatial domain and of the natural language instructions relative to it, or whether they merely overfit spurious signals in the dataset.

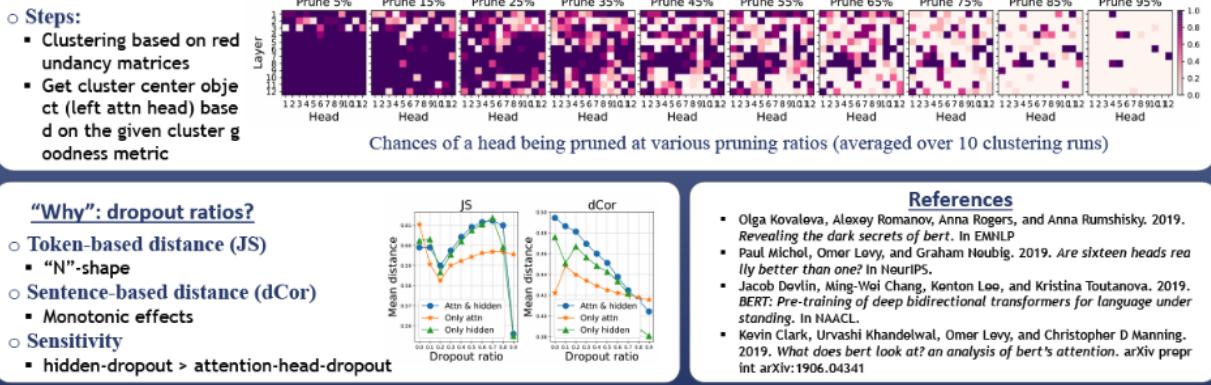
On attention redundancy: a comprehensive study

On Attention Redundancy: A Comprehensive Study

NAACL 2021

Yuchen Bian (yuchenbian@baidu.com), Jiaji Huang, Xingyu Cai, Jiahong Yuan, Kenneth Church Baidu Research





Towards interpreting and mitigating shortcut learning behavior of NLU models

Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU Models

Mengnan Du¹, Varun Manjunatha², Rajiv Jain², Ruchi Deshpande³, Franck Dernoncourt², Jiuxiang Gu², Tong Sun² and Xia Hu¹

¹Texas A&M University ²Adobe Research ³Adobe Document Cloud

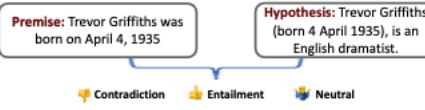


{dumengnan,xiahu}@tamu.edu
{vmanjuna,rajijain,rdeshp,franck.dernoncourt,jigu,tsun}@adobe.com



Shortcut Learning In BERT-based NLU Tasks

1. The NLU (natural language understanding) Task



2. The Explanation for BERT-based NLU models

neutral (1.00)	[CLS] on not nest as much as i'd like to i mean i've tend to stay pretty busy at my job and uh [SEP] my job wasn't se busy , i do that a lot more . [SEP]
entailment (0.67)	[CLS] equivalent to increasing national saving to 19 , [SEP] national savings are 15000 . [SEP]
contradiction (1.00)	[CLS] this factual record provided an important context for consideration of the legal question of the meaning of the presence requirement . [SEP] this record gave no context regarding the legal question . [SEP]
contradiction (0.68)	[CLS] hellenistic and roman periods [SEP] the hellenistic period . [SEP]
neutral (0.89)	[CLS] he thought phil number young husband would be the ones taking advantage of the argument about how cheating was hard to control . [SEP] husbands cheat on their wives . [SEP]
entailment (0.98)	[CLS] and i talked to someone about the uh uh education system i forgot exactly what the focus was on that one but that was fairly interesting and i've talked to somebody about credit card usage . [SEP] i talked to someone about the education system and credit card usage . [SEP]
entailment (0.99)	[CLS] the river plays a central role in all visits to paris . [SEP] the river is central to all visits to paris . [SEP]

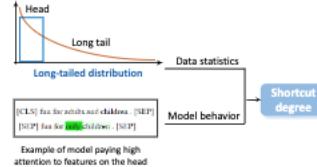
3. Our Shortcut Learning Observation

We provide explanations for BERT-based NLU model, finding that the model heavily relies on dataset biases as shortcuts for prediction:

- Paying attention to only hypothesis, rather than both premise and hypothesis
- Spurious statistics between simple unigrams and bigrams with labels
- Dropping from 87% accuracy on hold-out test set, to near random guess accuracy on adversarial test set

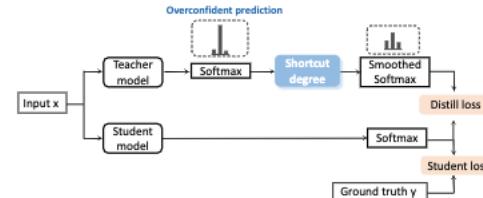
Long-Tailed Phenomenon for Interpreting Shortcut Learning

4. Long-Tailed Phenomenon by Comparing Dataset Statistics with Model Explanations



Self knowledge Distillation for Mitigating Shortcut Learning

5. Dis-encourage Model to Giving Overconfident Prediction for Shortcut Samples



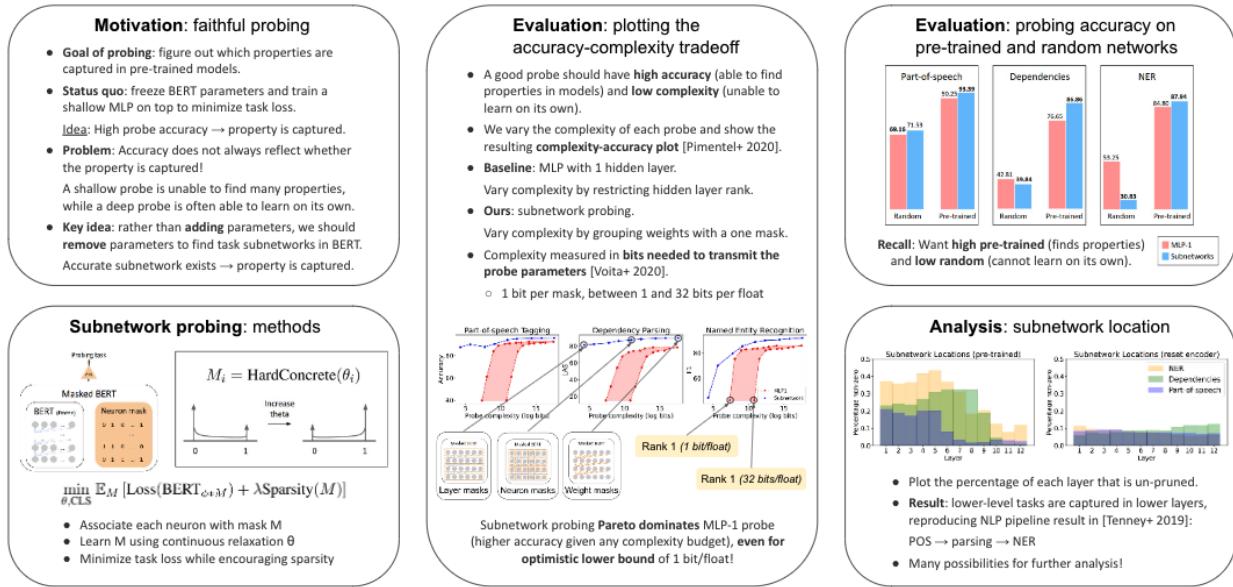
6. Our Model Relies Less on Shortcuts For Prediction

entailment (0.99)	the lot upon which it is being built had been vacant . [SEP] the lot had been vacant .
entailment (1.00)	the lot upon which it is being built had been vacant . [SEP] the lot had been vacant .

Low-complexity probing via finding sub-networks

Low-Complexity Probing via Finding Subnetworks

Steven Cao, Victor Sanh, and Alexander M. Rush



An empirical comparison of instance attribution methods for NLP

- Empirically study evaluating the degree to which different instance attribution agree with the importance of training samples.
- Simpler approximations can replace the complex ones.
 - The rankings are similar.
- Quality of similarity-based explanations are better than gradient-based methods.
 - Normalization to the gradient provides more consistency.
 - On HANS: Better attribution give higher influence to samples with high rate of overlap when mispredict entailment. → Similarity-based methods show higher lexical overlap.

Does BERT pretrained on clinical notes reveal sensitive data?

- Setup: Use EHR (MIMIC-III) to pretrain BERT.
- Masked language prediction: predict ICD-9 codes and MedCAT (disease / symptom tagger)
- Probing: removing the patient's name and simply encoding the condition to make a binary prediction yields similar (in fact, slightly better) performance
- Text generation: finding names (or names + conditions).

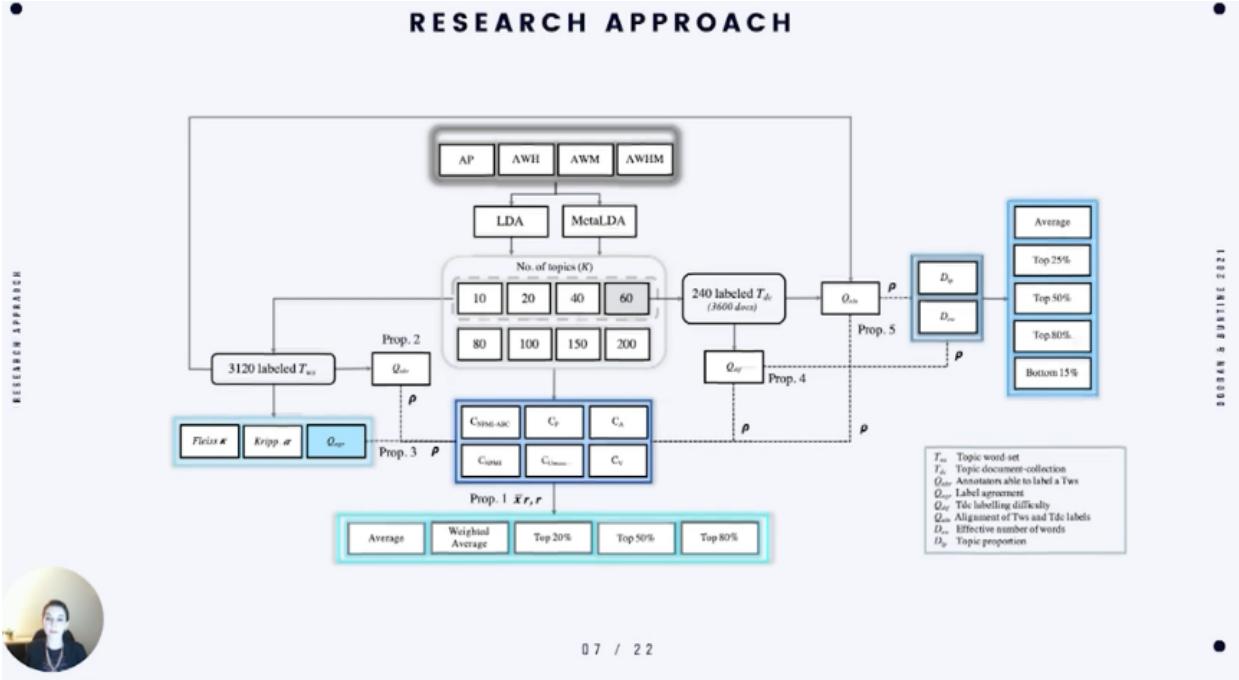
Interpretability analysis for NER to understand system predictions and how they can improve

- NER models learn mostly names. Context isn't learned even when the word is removed from the input. Build models that have access to part of the input.
- Is it possible to predict entity type solely from the context? To some extent.
- Is context aggregation optimal? No. Best possible aggregation by oracle vs aggregation by model. Oracle knows which of the models is correct.
- How can we utilize context better? Identify constraining contexts. Explore methods for better context clues aggregation.

Session 11B

Topic model or topic twaddle? Re-evaluating semantic interpretability measures

- Motivation: topic modelling for e.g., social media analysis has been increasing popular.
- Coherence scores: PMI, UMASS, C_A, NPMI, C_V, C_P
- Propositions:
 1. If coherence scores are robust, they should correlate.
 2. An interpretable topic is one that can be labelled.
 - Finding: No significant correlation between any coherence measure and Qnbr.
 3. An interpretable topic is one where there is high agreement on its label.
 - Qagr: agreement on the labels given to a topic between the four SME as a percentage.
 4. An interpretable topic is one where the document collection is easily labelled.
 - Qdiff: Labeling difficulty.
 - Qdiff and coherence, Qdiff and Dew: some relationships. No relationship between Qdiff and Dtp.
 5. An interpretable topic word-set is descriptive of its topic document collection.
 - Qalm: Rated alignment between Tws and Tdc.



Explaining neural network predictions on sentence pairs via learning word-group masks



NAACL 2021

Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks



IBM

Hanjie Chen¹, Song Feng², Jatin Ganhotra², Hui Wan², Chulaka Gunasekara², Sachindra Joshi², Yangfeng Ji¹
 hc9mx@virginia.edu¹ Department of Computer Science, University of Virginia² IBM Research AI

Abstract

Explaining neural network models is important for increasing their trustworthiness in real-world applications. Most existing methods generate post-hoc explanations for neural network models by identifying individual feature attributions or detecting interactions between adjacent features. However, for models with text pairs as inputs (e.g., paraphrase identification), existing methods are not sufficient to capture feature interactions between two texts and their simple extension of computing all word-pair interactions between two texts is computationally inefficient. In this work, we propose the Group Mask (GMASK) method to implicitly detect word correlations by grouping correlated words from the input text pair together and measure their contribution to the corresponding NLP tasks as a whole. The proposed method is evaluated with two different model architectures (decomposable attention model and BERT) across four datasets, including natural language inference and paraphrase identification tasks. Experiments show the effectiveness of GMASK in providing faithful explanations to these models.

Problem

- Explaining model predictions on text pairs

Natural Language Inference

Premise	An adult dressed in black holds a stick
Hypothesis	An adult is walking away, empty-handed
Label	Contradiction

- Identifying individual feature attributions

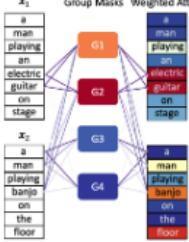
Limitations: ignoring feature interactions

- Detecting feature interactions

Limitations: computational complexity

Method

Group Mask (GMASK)



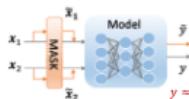
- Prediction: contradiction

- Distributing correlated words into a group (G1/G2/G3/G4)

- Learning word distributions and group importance

- Computing weighted word attributions (the weighted sum of group importance): "electric", "guitar" from x_1 , "banjo" from x_2

Generating local post-hoc explanations with word masks



Three properties of word masks

- correctly selecting important words for the model prediction
- removing as many irrelevant words as possible to keep the explanation concise
- selecting or masking out correlated words together from the sentence pair

Learning GMASK

- Objective $\min_{\phi, \psi} \mathcal{L}_{ce}(y, \hat{y}) - y_1(H(Z^U) + H(Z^L)) + y_2H(G)$

- Regularization on Z: ensure each group contains some words from both input sentences and avoid assigning a bunch of words into one group

- Regularization on G: ensure one or few groups have relatively large probabilities to be selected

- Optimization via sampling

$$W_{i,j} = \sum_{t=1}^T \delta(z_{i,t}) \delta(g_j)$$

Decompose word mask

$$\delta(a, b) = 1 \text{ when } a = b, \text{ and } 0 \text{ otherwise}$$

- Weighted word attributions

$$\theta_{i,j} = \sum_{i=1}^I \phi_{i,j}(i) \phi(i)$$

The expectation of $W_{i,j}$

Experiments

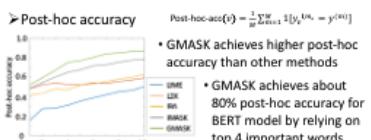
Setup

- Datasets: e-SNLI, Quora, QQP, MRPC
- Models: Decomposable attention model (DAttn), BERT
- Baselines: LIME [Ribeiro et al., 2016], L2X [Chen et al., 2018], IBA [Schulz et al., 2020], IMASK

$$\text{AOPC} = \frac{1}{(I-1)} \sum_{i=1}^{I-1} p(\hat{y}|x_i) - p(\hat{y}|x_{i+1})$$

✓ Higher AOPC is better

Models	Medians	e-SNLI	Quora	QQP	MRPC
DAttn	0.386	0.120	0.079	0.005	
L2X	0.354	0.137	0.104	0.089	
IBA	0.354	0.142	0.095	0.005	
IMASK	0.363	0.142	0.095	0.005	
LIME	0.221	0.155	0.110	0.062	
L2X	0.310	0.119	0.114	0.083	
IBA	0.292	0.232	0.139	0.130	
GMASK	0.319	0.309	0.181	0.200	



Degradation Test

- GMASK achieves higher degradation score

Models	Methods	e-SNLI	Quora	QQP	MRPC	$\text{deg-score} = \int_{-\infty}^{\infty} \mathbb{E}[\hat{y}(x_t) - \hat{y}^{(gt)}(x_t)] dx_t$
DAttn	LIME	0.502	0.070	0.090	1.367	
	L2X	0.453	0.110	0.197	2.775	
	IBA	0.455	0.157	0.254	2.077	
	IMASK	0.620	0.179	0.231	2.790	1.6
BERT	LIME	0.188	0.192	0.087	0.018	
	L2X	0.240	0.148	0.179	0.14	
	IBA	0.166	0.038	0.176	0.050	0.2
	IMASK	0.169	0.303	0.172	0.251	0.8
	GMASK	0.376	0.726	0.197	0.333	

Visualization of top four words

LIME, L2X, IBA, IMASK, GMASK

a man playing an electric guitar on stage, a man playing banjo on the floor

a man playing an electric guitar on stage, a man playing banjo on the floor

a man playing an electric guitar on stage, a man playing banjo on the floor

a man playing an electric guitar on stage, a man playing banjo on the floor

a man playing an electric guitar on stage, a man playing banjo on the floor

Discourse probing of pretrained language models

Tasks:

- (1) next sentence prediction (data: XSUM, Wikipedia),
- (2) sentence ordering. Shuffle 3-7 sentences, predict the right order (data: XSUM, wikipedia),
- (3) discourse connective: given two clauses, predict the connective (data: DisSent, CDTB, Potsdam),
- (4-6) RST nuclearity, relation, EDU segmentation (data: RST-DT, CDTB, Potsdam, RST-Spanish)

Findings:

- In understanding the discourse, BART's encoder and RoBERTa performed the best.
- Consistent pattern across different languages and model sizes: higher layers are in general better.

Learning to learn to be right for the right reasons

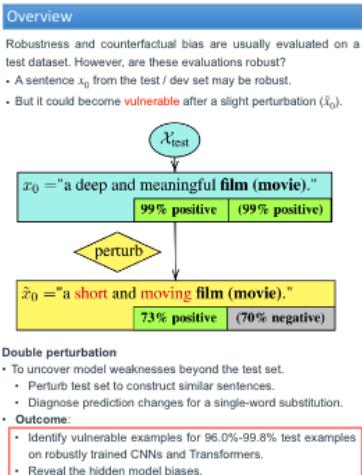
- Previously identified reasons for "superficial cues as bugs": loss discounting (Schuster+19), balancing token distribution (Kavumba+19), Adversarial filtering (Zellers+17), Adversarial training (Belinkov+19). Here → problem is with learning.
- Propose a method to learn to be "right for the right reasons"
- Evaluation: balanced-COPA, Commonsense Explanation.

Double perturbation: on the robustness of robustness and counterfactual bias evaluation

Double Perturbation: On the Robustness of Robustness and Counterfactual Bias Evaluation

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, Cho-Jui Hsieh
Department of Computer Science, UCLA

Code: github.com/chong-z/hlp-second-order-attack

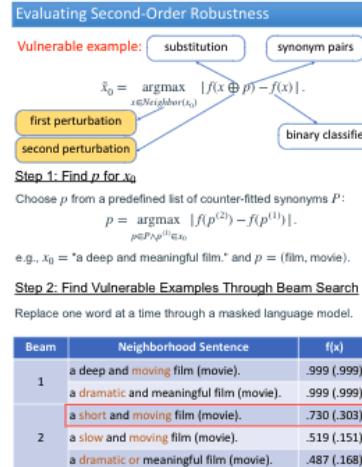


The Double Perturbation Framework

First Perturbation (non-label-preserving)
Perturb within a neighborhood. May affect the meaning.

large space small space

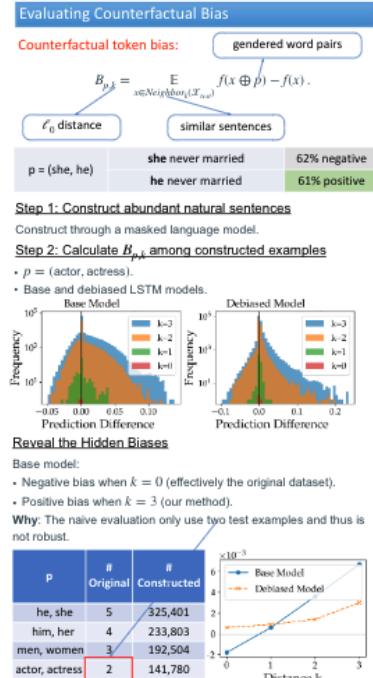
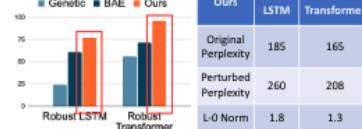
Second Perturbation (label-preserving)
Synonym substitution. Does not affect the meaning to human.



Experimental Results

High attack success rates by perturbing 1.3-1.8 words in average.

Attack Success Rate (larger is better)



- Motivation: it's possible to find similar but vulnerable sentences. Why? because the test set only consists of a small portion of possible natural sentences.
- Double perturbation framework:
 - First perturb the test set to construct abundant similar natural sentences. (Much larger space. Non-label preserving. Substitute words with a LM.)

- Then test if they are vulnerable. Label-preserving. Similar to existing attacks.
Substitute a single known-equivalent synonym.
- Successfully identify vulnerable examples for 77-99% of the test examples.
- Counterfactual bias: successfully reveal the hidden model bias not directly shown in the test set.

UniDrop: a simple yet effective technique to improve transformer without extra cost

- Analyze these types of dropout:
 - Feature dropout (on attention, activation, QKV, output)
 - Structure dropout (adopt LayerDrop)
 - Data dropout (given a sequence, with some probability keep the original sequence and do not apply data dropout)
- Theoretical analysis: these dropouts regularize different terms of the model. They can't be replaced by each other.
- Integrate these dropouts into UniDrop (how do they find the dropout rates of each? Hyperparameter tuning?)

Session: Interpretability bird-of-feather social

- Reliability testing for NLP systems <https://openreview.net/pdf?id=7ZL84tVIHZN>
- A diagnostic study of explainability techniques for text classification
<https://arxiv.org/abs/2009.13295>
- Evaluating RNN explanations <https://www.aclweb.org/anthology/W19-4813/>
- Towards faithfully interpretable NLP systems <https://arxiv.org/pdf/2004.03685.pdf>
- Randomizing BERT parameters and fine-tune on GLUE <https://text-machine-lab.github.io/blog/2020/bert-secrets/>
- Probing classifiers: promises, shortcomings, and alternatives
<https://arxiv.org/pdf/2102.12452.pdf>
- Evaluating attribution methods using white-box LSTMs <https://arxiv.org/abs/2010.08606>
- Quantifying attention flow in Transformers <https://arxiv.org/pdf/2005.00928.pdf>
- How does this interaction affect me? Interpretable attribution for feature interactions
<https://proceedings.neurips.cc/paper/2020/file/443dec3062d0286986e21dc0631734c9->

[Paper.pdf](#)

- Towards hierarchical importance attribution: explaining compositional semantics for neural sequence models <https://arxiv.org/abs/1911.06194>
- Probing with Shapley-value-based explanations as feature importance measures <http://proceedings.mlr.press/v119/kumar20e/kumar20e.pdf>

Papers in linguistic theory, psycholinguistics

Session 12E

On biasing Transformer attention towards monotonicity

 University of Zurich^{UZH}

On Biasing Transformer Attention Towards Monotonicity

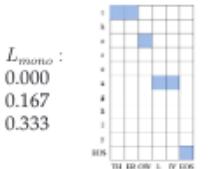
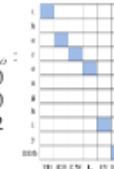
Annette Rios¹, Chantal Amrhein¹, Noëmi Aepli¹ and Rico Sennrich^{1,2}
¹Department of Computational Linguistics, University of Zurich
²School of Informatics, University of Edinburgh



Why Monotone Attention?

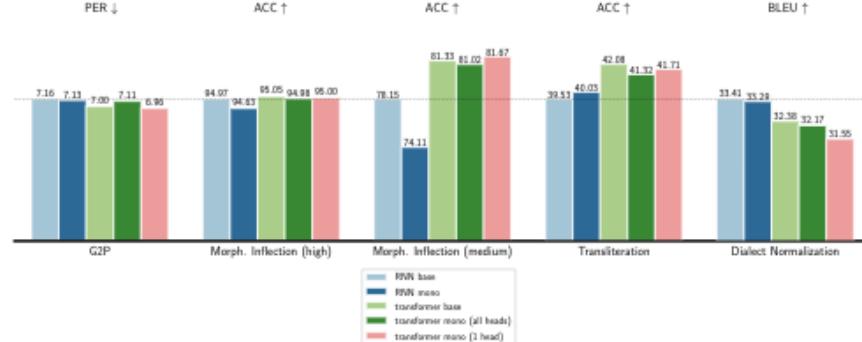
- beneficial for seq2seq tasks that are monotonic in nature (e.g. transliteration)
- enforced monotonicity on attention in RNNs beneficial in previous work
- transformers outperform RNNs even on highly monotonic tasks
- do transformers benefit from a bias towards monotonic attention?

Monotonicity Loss Function

Attention α :	Mean Attended Position \bar{a}_i :	Loss:
$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{ X } \exp(e_{ik})}$ $\delta = 0 \quad L_{mono} : 0.364$ $\delta = 0.5 \quad 0.530$ $\delta = 1 \quad 0.697$	$\bar{a}_i = \sum_{j=1}^{ x } \alpha_{ij} \cdot j$  $L_{mono} : 0.000$  0.167 0.333	$L_{mono} = \sum_{i=1}^{ Y -1} \max\left(\frac{\bar{a}_i - \bar{a}_{i+1} + \delta \frac{ X }{ Y }}{ X }, 0\right)$  0.000  0.000 0.152  0.000

Evaluation

- Transliteration (TR): News2015 shared task, 11 language pairs
- Grapheme-to-Phoneme Conversion (G2P): Nettalk and CMUDict, English
- Morphological Inflection (MI): CoNLL-Sigmorphon 2017 (high and medium), 51 languages
- Dialect Normalization (DN): Swiss German - German dataset



Code and data available: https://github.com/ZurichNLP/monotonicity_loss

Takeaway

- monotonicity in attention increased across all tasks and datasets
 - mixed results: improvements on some tasks
 - transformers: loss on all heads impairs ability to learn specialized functions
 - loss on subset of heads: beneficial on some tasks
- future work: explore usefulness of loss
 - where alignment is harder to learn
 - with more complex training schedule

Finding concept-specific biases in form-meaning associations



Are there cross-linguistic associations between the forms and meanings of words?

Arbitrariness of the Sign

Saussure claimed the association between word-forms and meanings is arbitrary.

- Example: Why is dog called cachorro in Portuguese?

Non-Arbitrariness of the Sign

Small but systematic patterns in these connections:

- Systematicity of the sign; Phonethemes; Iconicity.

Data - ASJP

- Basic vocabulary wordlists.
- 5189 languages! Almost ¾ of world's languages!
- 100 basic concepts: body parts, colour terms, ...

Non-Arbitrariness as MI

Operationalisation borrowed from our past selves (Pimentel et al. 2019):

$$MI(meaning, form) = H(form) - H(form | meaning)$$

Cross-entropy approximations:

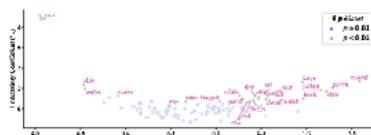
$$H(form) \leq H_0(form) \approx -\frac{1}{N} \sum \log p_{\theta}(form)$$

$$H(form | meaning) \leq -\frac{1}{N} \sum \log p_{\theta}(form | meaning)$$

Overall results

Macroarea	$H(W)$	$MI(W, V)$	$U(W V)$	Systematicity
Africa	3.77	0.011*	0.279%	
Americas	3.90	0.007	0.173%	
Eurasia	3.99	0.015†	0.376%	
Pacific	3.75	0.016‡	0.422%	
Average	3.85	0.012‡	0.312%	

Per concept



Conclusions

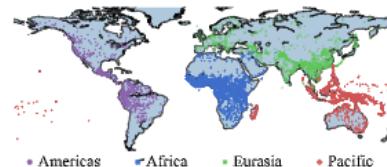
We propose a way to quantify cross-linguistic form–meaning biases

- We pointed out problems in moving from a within language analysis to a cross-linguistic setting and proposed solutions to them.

We find a set of concepts with particularly high mutual informations.

- These seem to drive most of the effect

In paper: extra per-language and concept--token association analysis!



Ab Antiquo: neuro proto-language reconstruction

- Can neural sequence models learn the regularities that govern historic sounds change in human languages?
- Train RNNs on two reconstruction tasks:

- Orthographic
- Phonetic
- Annotate dataset with 8000+ human-annotated entries in 6 Romance languages, derived from Wiktionary.
- Synthetic evaluation dataset
- Analysis of learned representation reveals the learning of phonologically meaningful representations without direct supervision.

How (non-)optimal is the lexicon?



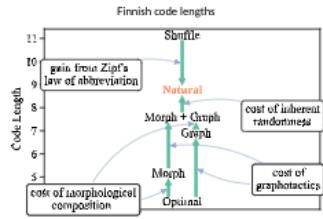
Language's Optimality

Researchers have talked about the "optimality" of language for a long time.

- Example: Zipf's law of abbreviation is taken as a sign of language efficiency.
- Counterexample: Short low-frequency words (wen) and long frequent words (happiness)

How far from optimal is language?

Can we measure the costs of specific linguistic constraints (e.g. morphology and graphotactics)?



Language as a Code

We take a coding-theoretic view of the lexicon.

- Meanings are messages;
- Words are codes;
- Listeners are receivers.

The expected code length for a language is:

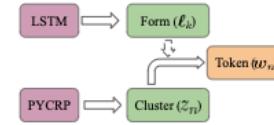
$$\text{cost}(\mathcal{C}) = \sum_{m \in \mathcal{M}} p(m) |\mathcal{C}(m)| \approx \frac{1}{N} \sum_{n=1}^N |\mathcal{C}(m_n)|$$

The Meaning Distribution

Assumption: one-to-one map of meanings to forms:
 $p(M = m_n) = p(W = w_n)$

We estimate this distribution using a neuralised version of Goldwater et al.'s (2011) two-stage model.

- Generator: Models wordforms
- Adaptor: Produces frequency distribution



Calculating Code Lengths

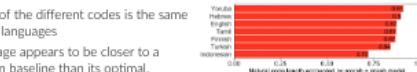
We estimate our codes as:

$$\begin{aligned} \text{cost} &\approx \frac{1}{N} \sum_{n=1}^N |\mathbf{w}_n|, & \text{cost} &\lesssim \frac{1}{N} \sum_{n=1}^N \left[\log(\Sigma) \frac{1}{p_0(\mathbf{w}_n)} \right], & \text{cost} &\approx \frac{1}{N} \sum_{n=1}^N |\mathbf{w}'_n|, \\ \text{natural code} && \text{optimal code} && \text{graphotactic code} \\ \text{cost} &\lesssim \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^{L_n} \left[\log(\Sigma) \frac{1}{p(\mathbf{u}_{n,j})} \right], & \text{cost} &\approx \frac{1}{N} \sum_{n=1}^N \sum_{j=1}^{L_n} |\mathbf{u}'_{n,j}|, \\ \text{morphology code} && && \text{morpho+graph code} \end{aligned}$$

* We rely on Morfessor to get individual "morphemes".

* We sample wordforms from an LSTM to get our graphotactics code.

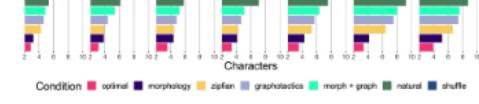
Results



* Order of the different codes is the same across languages

* Language appears to be closer to a random baseline than its optimal.

* Morphology and graphotactics account for most of the natural code length



Linguistic complexity loss in text-based therapy

Linguistic Complexity Loss in Text-Based Therapy

Jason Wei, Kelly Finn, Emma Templeton, Thalia Wheatley, and Soroush Vosoughi

NAACL 2021

Complexity Loss Paradox (Goldberger, 1967): individual sufficiency from disease

1997): Individuals suffering from disease exhibit surprisingly predictable behavioral dynamics.

- Or, "Animals lose complex behavior under stress."

Observed in...

- Diving patterns in penguins
 - Cyclic oscillations in white blood cell counts in leukemia patients

Our paper's question:

- What linguistic complexity patterns in the language of clients and therapists during therapy reflect client mental health?

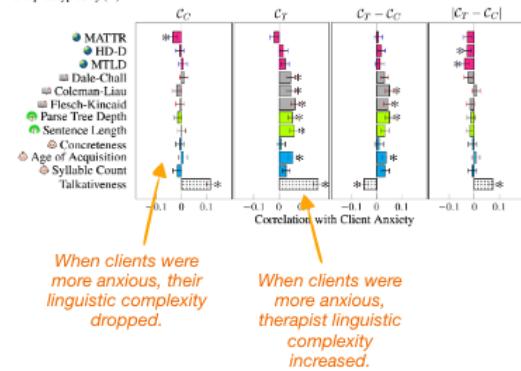
Talkspace Therapy Dataset

Table 1: Descriptive statistics for Talkspace online therapy conversations dataset. † indicates mean.

	Dataset		
	Exploratory	Confirmatory	
Messages	2.6 million	0.7 million	> 3 million messages
Survey responses	24,287	6,150	
Clients	5,736	1,434	
Therapists	1,608	889	
#Survey responses / client	4.23	4.29	
#Client text (words) / survey	1259	1295	~ 4 survey responses per client indicating anxiety over time!
#Therapist text (words) / survey	796	804	
Median survey score (0-21)	8	8	
Median time between surveys	21 days	21 days	

Linguistic Complexity Loss

Figure 1. Linguistic complexity measures correlate with client anxiety (Φ indicates significance at $p < 0.001$ for both the exploratory and confirmatory datasets). We show correlations ($\pm 95\%$ confidence intervals) on the exploratory dataset for language and client complexity of clients (C_{C1}), therapists (C_{T1}), therapist and client difference ($C_{C2} - C_{T1}$), and absolute therapist and client difference ($|C_{C2} - C_{T1}|$). Each complexity measure was entered into its own one-line mixed model. We group complexity measures into lexical diversity (●), syntax (○), readability (■), and prototypicality (▲).



Variation in Complexity Loss

Table 3: \pm indicates how much individuals varied linguistic complexity among their own messages compared with a random sample from the population. We show average \pm for within-individual standard deviation σ and range Δ , for clients (C) and therapists (T). * indicates significance at $p < 0.001$ for both exploratory and confirmatory datasets.

	Standard Deviation (σ)			Range (Δ)					
	σ_{P}	σ_{S}	σ_{PS}	σ_{P}^2	σ_{S}^2	σ_{PS}^2			
MATTR	-0.6	-0.33	-0.47	0.36 ^a	-0.35	-0.29	-2.17	0.0398	
HD-D	-0.3	-0.22	0.53	0.2936	-0.3	-0.28	-0.63	0.0482	
MTLD	-0.36	-0.35	-0.06	0.9581	-0.35	-0.34	-0.17	0.8695	
Dale-Chall	-0.44	-0.65	6.43 ^a	<0.0001	-0.45	-0.41	0.51	0.91	0.0063
Coleman-Liu	-0.68	-0.74	1.91	0.0258	-0.68	-0.61	-1.84	0.15	0.0063
Flesch-Kincaid	-0.46	-0.93	15.08 ^a	<0.0001	-0.47	-0.76	-13.33 ^a	<0.0001	
Pearce Depth	-0.77	-0.89	4.12 ^a	<0.0001	-0.74	-0.73	-0.48	0.6324	
Sentence Length	-0.44	-0.97	17.69 ^a	<0.0001	-0.45	-0.34	-15.38 ^a	<0.0001	
Concreteness	-0.49	-0.36	-0.45 ^a	<0.0001	-0.48	-0.42	-0.73	0.57 ^a	<0.0001
Age of Acquisition	-0.44	-0.09	-0.44 ^a	<0.0001	-0.41	-0.47	-0.36	0.53 ^a	<0.0001
Lexical Density	-0.31	-0.31	0.41	0.01	-0.31	-0.31	-0.01	0.62 ^a	<0.0001
Traitlessness	-0.31	-0.36	1.94 ^a	<0.0001	-0.3	-0.58	11.88 ^a	<0.0001	

Both clients and therapists had "unique voices" in terms of linguistic complexity.

Word complexity is in the eye of the beholder

- Task: Complex word identification
 - Claim: (Current CWI systems follow "one-size-fits-all" approach) CWI should be different, depending on the audience (e.g., native vs. non-native)
 - Release a CWI dataset, annotated by readers with different backgrounds.

Language in a (search) box: grounding language learning in real-world human-machine interaction



Language in a (Search) Box: Grounding Language Learning in Real-World Human-Machine Interaction

Federico Bianchi, Ciro Greco, Jacopo Tagliabue



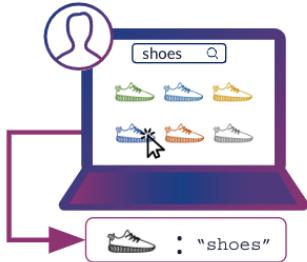
Meaning is grounded in objects

Language is used to refer to extra-linguistic entities: linguistic meaning can be represented as a mapping between words and the things they refer to.



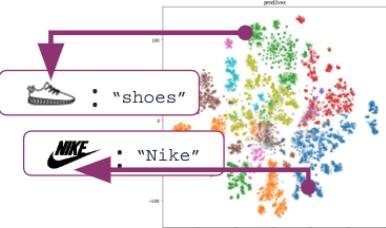
Using IR to learn a grounded semantics for noun phrases end-to-end

- Fully learnable object-based semantics in the context of a product search engine: object domain, denotation and compositionality are learned without tagging;
- able to support zero-shot generalization like symbolic approaches.



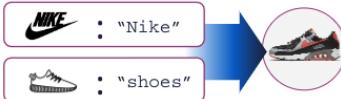
Lexical denotation as DeepSets

1. A dense objectual domain is learned from clickstream data with `prod2vec`.
2. The meaning of "shoes" is a **DeepSet** of objects (average pooling of product embeddings), as mapped through the User-Engine dynamics.



Deep compositionality

The meaning of "Nike shoes" depends functionally on the meaning of its constituents - $\text{DeepSet} \times \text{DeepSet} \rightarrow \text{DeepSet}$:



We test both Additive Compositional Model (ADM) and Matrix Compositional Model (MDM) as our composition strategies.

Experiments

1. **Leave-one-brand-out (LOBO)**: we train models over "brand + object" queries but we exclude a specific brand; in the test phase, we predict the DeepSet for a seen object and an unseen brand (e.g. "Nike shoes", where "Nike" was not in the training set).
2. **Zero-shot (ZT)**, we train models over two-terms phrases (e.g. "Nike shoes", "soccer shoes", "men shoes") and test generalization on unseen, more complex NPs (e.g. "Nike basketball shoes").

Intra-textual baselines

1. **BERT (UM)**: we extract the 768-dimensional representation from the [CLS] embedding and learn a linear projection to the product-space.
2. **W2VEC (W2V)**: we learn a compositional function that concatenates **DeepSets**, projects them to 24 dimensions, passes them through a Rectified Linear Unit, and finally projects them to the product space.

Results

MDM and ADM significantly outperform UM and W2V on both tasks.

nDCG	MDM	UM	W2V
LOBO	0.299	0.002	0.009
ZT	0.098	0.032	0.006

TAKE AWAY: a dense object domain, encoding properties in the topology of the space, can underpin compositionality on a discrete level – symbolic-like inference emerges from a fully dense domain.

Papers in Computational Social Science

Session 6E

The structure of online social networks modulates the rate of lexical change



The structure of online social networks modulates the rate of lexical change

Jian Zhu and David Jurgens
University of Michigan



Research goal

In sociolinguistics, one structural factor that has long been recognized as influencing lexical changes is the language community's social network.

In this study, we examine how network structures affect lexical change in online communities. Specifically,

- How does network structure contribute to the introduction of new words to online communities (**innovation**)?
- How do structural properties affect the survival of these newly introduced words (**retention**)?
- Does the increased inter-connectedness cause online communities to adopt a similar set of new words (**levelling**)?

The Reddit Corpus

We selected the top 4420 subreddits based on their overall size from 2005 to October 2018.

Intra-community networks

- undirected and unweighted graphs.
- Each user is represented as a node
- An edge exists between users if these two users have interacted in close proximity.

Community networks

- a weighted and undirected network with the edge weights set to the numbers of shared users.
- A community is represented as a node in the graph.
- Two communities are determined to be connected if they share active users.

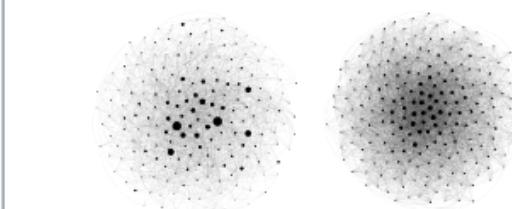
Internet neologisms

We obtained 80071 Internet neologisms Internet slangs from two online dictionary sources, NoSlang.com and Urban Dictionary.

Frequency	Neologisms
Frequent	lol, /r, kinda, bitcoin, idk, lmao, tbh, tl;dr, a lot, /s, omg, lwl, bahaha, iirc
Infrequent	thugmonster, blain, solk, f'ang, yobbish, ferranti, sonse, vampy

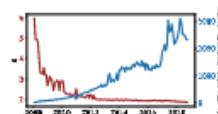
Selected Findings

After controlling for size, the network with **higher average degree** (more inner-connections) (right: r/F13thegame) tends to develop **more lexical innovations** than the one with **lower average degree** (left: r/MassEffects), which is not consistent with the classic **weak tie model of change**.



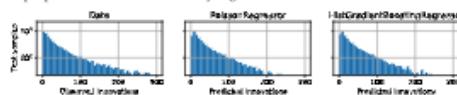
Levelling

- Levelling refers to the gradual replacement of localized linguistic features by mainstream linguistic features.
- Online communities under investigation seem to go against the levelling trend observed in offline networks.
- The number of community specific words grew rapidly (**decreased α** below) despite increased inter-community connectedness (**increased average community degree** below)
- Segregation in topics and interests naturally brings in more community specific words.



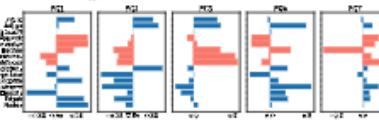
Lexical innovation

- We ran regression analysis on the count of innovation for each monthly subreddit using structural features of both inter- and intra-community network.
- Both Poisson regression model and Gradient Boosting Tree model can predict lexical innovation above the random baseline.
- Structural properties can account for many regularities in the creation of lexical innovations.



Lexical survival

- Here, we test whether **network features** systematically affect the **survival of words** (durations of survival) in online communities using survival analysis.
- The figure below shows the five most important principal components for predicting word survival.
- **A large overall size** tends to preserve neologisms, as large communities provide a basic threshold population for words to be used.
- Global network features such as **high average degree**, **high network centrality** and **strong connectedness** also contribute to neologism survival.



Conclusions

- Conclusion
- The **overall network size** is the most prominent factor in lexical innovation and survival, as large communities provide the base population to create and use neologisms.
- **Dense edges between users, the lack of separate local clusters, and rich external connections** also promote both lexical innovation and survival.
- Lexical change process in online social networks may be similar to other information spread processes.
- Our quantitative analysis also suggests a different leveling process in online communities with implications for sociolinguistic theories.

Acknowledgements

Nic Shanks, Professor Patrick Reiher, Professor Paul Fletcher, John Mendelsohn, Jason Pei, Danya Tan, Alfonso Labraak and anonymous reviewers for their comments on earlier versions of this draft. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1234222.

Framing unpacked: a semi-supervised interpretable multi-view model of media frames

Framing Unpacked: A Semi-Supervised Interpretable Multi-View Model of Media Frames

Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, Andrew Turpin, Lea Frermann



Introduction

Framing: selecting some facts over others, and make certain perspectives more salient.

- Equivalence framing**: expressing the same semantics in different forms.
- Emphasis framing**: presenting selective facts and aspects.
- Story framing**: Using narrative structures to convey information.

Previous work -> Emphasis framing
Our model (FRISS) -> Emphasis framing, Story framing

Intuition

- Explain in terms of local framing signals
- Learn frame-specific latent representations

The Obama administration's decision to move forward with a legal challenge to Arizona's stringent illegal immigration law will almost certainly cleave the issue on the campaign trail this fall. The Arizona measure, which was signed into law by Gov. Jan Brewer (R) in April, is a major political touchstone of prime importance to Hispanics, the fastest growing demographic group in the country and a coveted electoral prize for both parties. Democratic strategists see the Arizona law as a key moment in the ongoing battle to win the loyalty of Hispanic voters. They believe that it will have a similar chilling effect for Republicans with Latinos as the passage of California's Proposition 187 did in the 1990s. Republicans, on the other hand, believe the Democrats are likely to be up with conservatives on the immigration issue. They said the Obama administration's aggressive approach to fighting the Arizona law is yet more evidence of that out-of-touchness. In that vein, nearly 190 House Republicans ~~were asked~~ in August General Eric Holder on Tuesday, describing the legal challenge as "the height of irresponsibility and arrogance." Polling on the Arizona law specifically finds Republicans' favor, although broader data suggests a public deeply divided on immigration. In the latest Washington Post/ABC poll, 58 percent expressed support for the Arizona law – including 42 percent who were strongly supportive – while 41 percent opposed it.

■ Political ■ Legality ■ Public Sentiment

- The true frame label of article is "Political".
- Our model can detect local framing signals related to different frames

The Proposed Model (FRISS)

Unsupervised Module

Supervised Module

Experiment 2: Benefit of Unlabelled Data

Experiment 3: Qualitative Evaluation

ARGO	ARG1	Predicate
Trump, house republican, Obama, democrat, senate		
supreme court, justice, federal judge, court		
organizer activist, protester, demonstrator, marcher		
amendment, reform, legislation, voter, senate, bill		
political asylum, asylum, lawsuit, status, case		
rally, marcher, march, protest, movement, crowd		
veto, defeat, vote, win, introduce, endorse, elect		
sue, uphold, entitle, appeal, shall, violate, file		
chant, march, protest, rally, wave, gather, organize		
Political Legality Public Sentiment		

V. Conclusion

- Developed a novel semi-supervised frame classification model.
- Leveraging unlabelled data our model can improve document level frame prediction.
- Latent multi-view representations add interpretability and nuance to predicted document frames through local semantic roles.

Code: <https://github.com/shinyemimalef/FRISS>

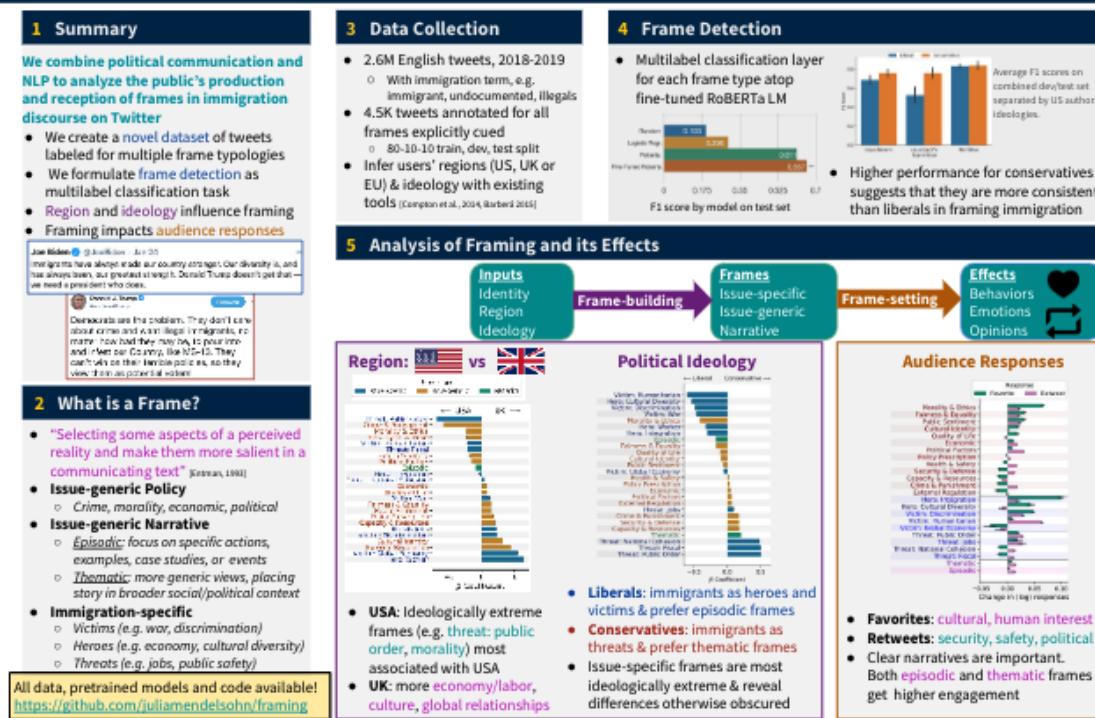
Modeling framing in immigration discourse on social media

NAACL 2021 papers

29



Modeling Framing in Immigration Discourse on Social Media



Automatic classification of neutralization techniques in the narrative of climate change scepticism

- Introduce the NT multilabel classification task for climate change scepticism
 - Labels: condemn (used to blame the alarmist greens), D-responsibility (used to highlight global warming being a natural cycle)

WikiTalkEdit: a dataset for modeling editors' behaviors on Wikipedia

- Discussions on Wikipedia talk pages could help persuade editor behaviors
- An example of exploratory analysis: x: positive emotional change. y: editorial change.

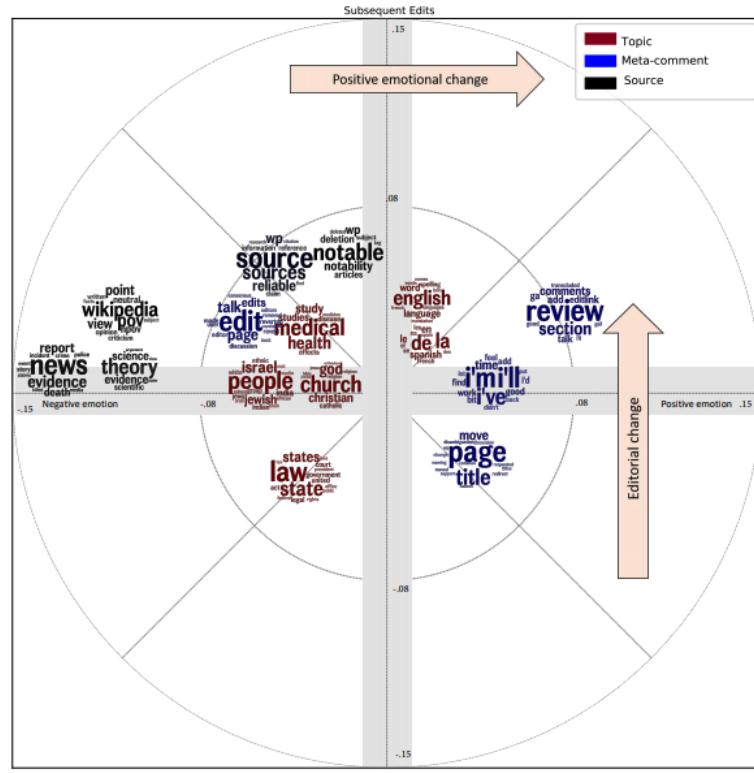


Figure 2: LDA topics correlated with emotional and editorial change. Topics are colored according to their theme; word size is proportional to word weight in the topic.

- More descriptions about dataset in [paper](#).

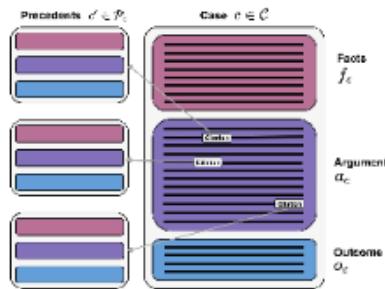
Session 7A

What about the precedent: an information-theoretic analysis of common law

What About the Precedent: An Information-Theoretic Analysis of Common Law

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, Simone Teufel

Halsbury (1907): Goodhart (1930):
Arguments as ratio Facts as ratio



Fact	<i>The applicants, D.P. and J.C., who are sister and brother, are United Kingdom nationals, born in 1964 and 1967 and living in London and Nottingham, respectively...</i>
Argument	<i>Article 2 of the Convention provides, in its first sentence: "1. Everyone's right to life shall be protected by law; ..." 46. The applicants complain that the authorities failed to protect the life of their son and were responsible for his death...</i>

$$\text{MI}(O; H | F) = H(O | F) - H(O | H, F)$$

$$\text{MI}(O; G | F) = H(O | F) - H(O | G, F)$$

Arguments as ratio

Where random variables O , H and F represent:
 O : Outcome of the case at hand
 H : Arguments and Outcomes of the precedent cases + Facts of the case at hand
 F : Facts of the case at hand



Where random variables O , G and F represent:
 O : Outcome of the case at hand
 G : Facts and Outcomes of the precedent cases + Facts of the case at hand
 F : Facts of the case at hand



Facts as ratio

Characterizing English variation across social media communities with BERT



Characterizing English Variation across Social Media Communities with BERT

Li Lucy & David Bamman
University of California, Berkeley

Summary

We measure semantic variation at scale across hundreds of Reddit communities, using an efficient method involving BERT embeddings.

- We validate this method using standard benchmarks and in-domain, user-created glossaries.
- We pair this type of variation with a more traditional approach of identifying distinctive word types.
- **Communities with distinctive language are medium-sized, and their loyal and highly engaged users interact in dense networks.**

Our dataset has comments (1.4+ billion tokens) from 474 subreddits written during May-June 2019.

Methods for Identifying Community-Specific Language

Word Type

Past work on online language norms has focused on lexical choice. We experiment with several methods for finding salient and distinctive words in each community: PMI, NPMI, tf-idf, TextRank, and Jensen-Shannon divergence.

For example, for word i in subreddit s , its NPMI is:

$$T_i(s) = \log \frac{P(s, i)}{P(s) P(i)}$$

Examples of words with high NPMI in a subreddit:



Word Senses

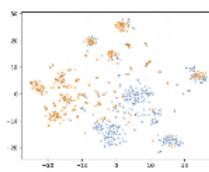
Communities can also systematically use the same word to mean different things.

1. Our word sense induction (WSI) method runs k-means directly on BERT embeddings.
2. Amrami & Goldberg 2019's SOTA WSI method clusters word substitutes, which BERT predicts for a masked target word. The SOTA model performs better on SemEval benchmarks, but our method is 48x faster to scale on Reddit data and shows similar performance within that domain.

Our metric for community-specific senses

The sense NPMI of a word = the NPMI of their most common sense in a subreddit. Examples of words with high sense NPMI:

subreddit	word	subreddit example	other sense example
r/libertarian	nap	"The nap is just a social contract."	"Move bedtime earlier to compensate for no nap..."
r/90dayfiance	nickel	"Nickel really believes that Azan loves her."	"...raise burrito prices by a nickel per month..."



r/elitedangerous: The [MASK] is a good multipurpose ship and a spectacular ship for grinding through missions.
→ Trident, brig, pilot, mermaid, viper, pirates

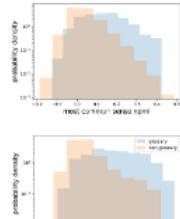
r/cscareerquestions: No I've used [MASK]. HTML, CSS, Javascript, node, flask
→ slack, oracle, apache, bat, framework, windows

Top: t-SNE of BERT embeddings for python in two subreddits. Bottom: BERT substitutes for python.

Glossary Analysis

Are words determined by users as important to their communities also emphasized by methods for identifying community-specific language? Yes.

- We collected **57 user-created glossaries** containing 2800+ words from subreddit wiki pages for in-domain validation
- We examined the **percentage of glossary words in the 98th percentile of scored words** and the **mean reciprocal rank** of the highest scored glossary word.
 - NPMI for word types (bottom right) was the best metric for capturing the concept of community-specific language.
 - Sense NPMI with BERT embeddings (top right) and sense NPMI with BERT substitutes behaved similarly. Our later analyses focus on the former.
- **NPMI for finding distinctive word types and NPMI for word senses are complementary to each other**, where only 21 glossary words are in the 98th percentile of both.

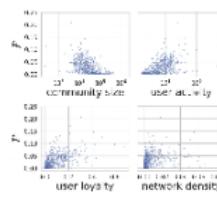


The distribution of scores for glossary words (blue) is higher than that for non-glossary words (orange)

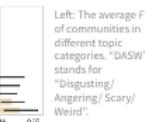
Community-level Attributes & Variation

What kinds of communities have very distinctive language?

- Smaller communities
- More active communities, where activity = avg # of comments per user
- Communities with **more loyal** users, where loyalty = over 50% of a user's comments are in that community. Loyal users also tend to use words that are part of the community's sociocultural style more often.
- Communities with **more dense direct-reply networks**



Left: the relationship between each community-level attribute and F , or the fraction of words in a community in the 98th percentile of type NPMI or embedding sense NPMI. Each point is a subreddit.
Communities with distinctive word types also tend to have distinctive word meanings (Spearman corr = 0.7855, $p < 0.001$).



Left: The average F of communities in different topic categories. "DASW" stands for "Disgusting/Scary/Weird".

Communities with topics related to Video Games, TV, Sports, Hobbies/Occupations, and Technology tend to have more community-specific language.

However, in a regression analysis, though topic has a higher effect on community-specific language, user activity and loyalty each had more of an effect. This suggests that **who is involved in a community may matter more than what they discuss**.

Overall, our results confirm several sociolinguistic hypotheses related to the behavior of users and their use of community-specific language.

Session 7B (Green NLP)

It's not just size that matters: small LMs are also few-shot learners

IT'S NOT JUST SIZE THAT MATTERS SMALL LANGUAGE MODELS ARE ALSO FEW-SHOT LEARNERS

Timo Schick and Hinrich Schütze
CIS, LMU Munich, Germany | schick@cis.lmu.de

1 What Problem Are We Trying To Solve?

Few-Shot Learning

Learning tasks **only from a few examples** is a key challenge for NLP. To illustrate this, try guessing the **correct output** for the last input:

This was the best pizza I've ever had! 0
 You can get better sushi down the road for half the price. 1
 Salmon nigiri was bad. Not worth what they're asking. 1
 Excellent pizza! Slices are fantastic, prices are reasonable. ?

2 How Do We Approach This Problem?

Pattern-Exploiting Training

Pattern-Exploiting Training (PET) facilitates few-shot learning by providing a masked language model M with **task descriptions**. This requires:

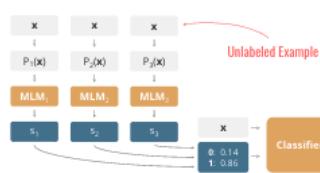
- A **pattern P** that converts each input into a cloze question
- A **verbalizer v** that expresses each output in natural language



3 So, How Exactly Does PET Work?

Combining Task Descriptions

As finding a single task description that works well can be challenging, PET enables the **combination of multiple pattern-verbalizer-pairs**:



PET with Multiple Masks

INFERENCE

$p_i(\text{econom}) < p_i(\text{ics})$

$p_i(\text{ics}) : \text{Bat-winged dinosaurs were clumsy fliers.}$
 $p_j(\text{econom})$

$$p(\text{I}) = p_i(\text{ics}) \cdot p_j(\text{econom})$$

TRAINING

$p_{\text{MASK1}} \approx p_{\text{MASK2}}$

$$p(\text{I}) = p_{\text{MASK1}}(\text{science}) \cdot p_{\text{MASK2}}(\text{econom})$$

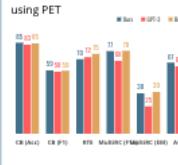
4 And How Well Does PET Work?

Results on SuperGLUE

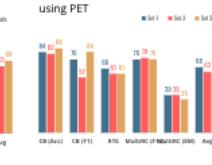
with ALBERT-xxlarge-v2 for 32 examples



Different Patterns using PET

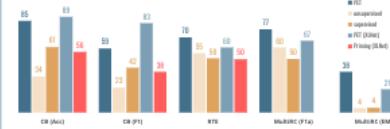


Different Training Examples using PET



Different Few-Shot Methods

on selected SuperGLUE tasks



This work was funded by the European Research Council (ERC StG 757516).

Get the **Paper** and **Code**: <http://timoschick.com/naacl2021/>

Static embeddings as efficient knowledge bases?

Static Embeddings as Efficient Knowledge Bases?

Philipp Dufter*, Nora Kassner*, Hinrich Schütze

Center for Information and Language Processing (CIS) LMU Munich, Germany

(philipp,kassner)@cis.lmu.de

Motivation

- Probing factual knowledge captured by Pre-trained Language Model:
“The capital of France is [MASK].” (LAMA)
- ⇒ Pretrained Language Models encode to some extend **factual knowledge**
- Obtain insights in the **underlying mechanism**
- Compare with **static embeddings**

Comparison

Model	Vocab. Size	p1	
		LAMA	LAMA-UHN
BERT	30k	39.6	30.7
mBERT	110k	36.3	27.4
	30k	16.4	5.8
	120k	34.3	25.0
fastText	500k	39.9	31.8
	1000k	41.2	33.4

Conclusions

1. Static embeddings **competitive and cheap**
2. BERT great at **composing** representations from subwords
3. BERT considers **relation** information
4. **Underlying mechanism** in BERT not more effective than NN-matching
5. Simple and “green” worth to be considered

Typed Querying

Contextualized embeddings:

$\arg \max_{c \in \mathcal{C}} p(c|t)$
 \mathcal{C} candidate set {"Paris", "London", "Berlin", ...}
 t template ("[X] is the capital of [MASK].")

Static embeddings:

Nearest neighbor matching
 $\arg \max_{c \in \mathcal{C}} \text{cosine-sim}(\bar{e}_q, \bar{e}_c)$,
 $\bar{e}_c = \frac{1}{k} \sum_{i=1}^k e_{t_i}$ mean pooled representation
 c gets wordpiece tokenized into t_1, \dots, t_k
disregards relation information

Advantages:

1. Understands type constraints
2. Focuses on the knowledge intensive part
3. Comparability across contextualized/static emb.

Evaluation

We compute precision at one (p1) for each relation, i.e., $1/|T| \sum_{t \in T} \mathbb{1}\{\hat{t}_{\text{object}} = t_{\text{object}}\}$ where T is the set of all triples and t_{object} is the object predicted.

Dataset: TREx with 41 relations; 10 languages.

Multilingual Results

Model	Vocab. Size	p1								
		AR	DE	ES	FI	HE	JA	KO	TH	TR
Oracle		21.9	22.3	21.6	21.3	22.9	21.3	21.7	23.7	23.5
mBERT	110k	17.2	31.5	33.6	20.6	17.5	15.1	18.9	13.5	33.8
	30k	20.8	16.2	17.1	16.7	21.4	14.6	17.3	21.3	22.1
	120k	27.9	25.2	31.0	24.2	28.3	22.4	28.2	28.0	33.2
fastText	500k	31.7	32.5	36.6	30.9	33.7	27.0	31.5	31.8	36.1
	1000k	31.3	33.6	36.5	31.8	33.9	27.2	29.8	30.5	36.6

⇒ with **large vocab** fastText becomes competitive

Resource Consumption

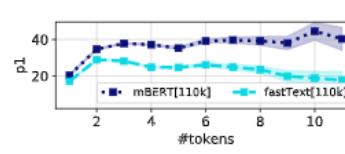
- Static embeddings much **cheaper** to compute
- 0.3% of carbon emissions of BERT
- Only **CPU** required

Prediction Diversity

Model	Vocabulary Size	p1-mf	Entropy	#Distinct pred.
BERT	30k	35.7	6.48	85
fastText	1000k	42.5	7.32	119

⇒ fastText predictions are **more diverse**

Contextualization



- ⇒ mean pooling worse for multitone subjects
⇒ **trade-off**: learning good composition function with small vocabulary vs. large number of atomic representations

This work was supported by the European Research Council (# 740516) and the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content. The first author was supported by the Bavarian research institute for digital transformation (bidt) through their fellowship program. We thank Yanai Elazar and the anonymous reviewers for valuable comments.

Presented at NAACL 2021, Online. * Equal Contribution

Session 11A (ethics)

On the impact of random seeds on the fairness of clinical classifiers

On the Impact of Random Seeds on the Fairness of Clinical Classifiers



We investigate the impact of **random seeds** on the **fairness** of fine-tuned classifiers with respect to demographic characteristics such as gender and ethnicity.

fairness: mean differences in model performance across demographic subgroups

DATA AND METHODS

- We used **MIMIC-III** to develop classifiers for:
 - In-hospital Mortality Prediction
 - Phenotype Classification
- We used a pre-trained ClinicalBert model to induce **fine-tuned clinical classifiers** from clinical notes
- We sampled $k=1000$ pairs of random seeds from $\mathcal{U}(0,1000)$. For each seed pair, we compared the overall model performance (AUC) with performance for each subgroup (ΔAUC)

$$\Delta AUC = AUC_{\text{subgroup}} - AUC_{\text{overall}}$$

MIMIC III

EHR from 40K patients admitted to the ICU of the Beth Israel Deaconess Medical Center between 2001-2012

- Vitals, Labs, Clinical notes
- Protected Attributes (e.g., demographics)

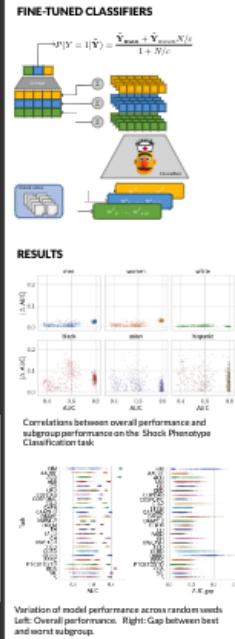
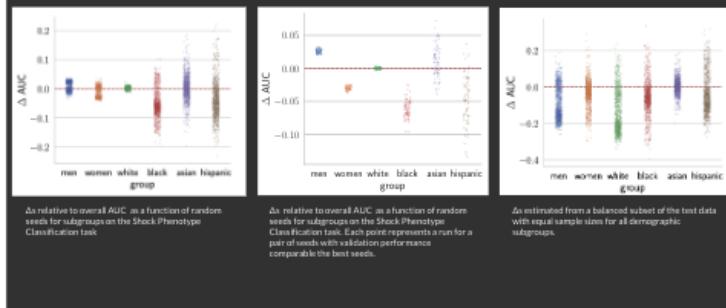


Silvio Amir, Jan-Willem van de Meent, Byron C. Wallace
 {s.amir, b.wallace, j.vandemeent}@northeastern.edu

Random seeds significantly impact the performance and fairness of clinical classifiers **on MIMIC-III**

Studies of **algorithmic fairness** should account for:

- model variability due to choice of **random seeds**
- variance due to **small sample sizes**



Dynamically disentangling social bias from task-oriented representations with adversarial attack

- Social bias, protected attributes.
- Related work e.g., INLP. They are mostly post-processing steps → static.
- Use adversarial training to achieve debiasing.
- Baselines: original classifier, INLP, random noise.
- Evaluation metric: TPR gap (debiasing task), sentiment (main task).

An empirical investigation of bias in the multimodal analysis of financial earnings calls

AN EMPIRICAL INVESTIGATION OF BIAS IN THE MULTIMODAL ANALYSIS OF FINANCIAL EARNINGS CALLS

RAMIT SAWHNEY, ARSHIYA AGGARWAL, RAJIV RATN SHAH

KEY IDEAS

- Verbal and vocal cues in financial disclosures improve volatility prediction
- Potential bias that models may learn from the speech signals of company executives
- Analysis of gender bias as error disparity in volatility prediction from the multimodal CEO speeches in financial earnings calls



MOTIVATION

- Multimodal approaches improve risk prediction
- Audio features vary across genders
- Female underrepresentation in the financial realm
- Imperative to understand and quantify bias in financial multimodal language data

~5%
FORTUNE
500 CEOs
ARE
WOMEN

ONWARDS

- Expand our study to different sensitive attributes like region, ethnicity, accent etc.
- Augmentation techniques and cross domain adaptation to improve female performance

METHODS

- Use publicly available audio recordings and transcripts of CEO speeches.
- Map speakers to self-reported gender using Reuters, Crunchbase, Wikidata.
- Quantify gender bias as error disparity in state-of-the-art multimodal deep regression model (MDRM).
- Perform statistical tests on the male-female audio features
- Use different male:female ratios for train-set to get deeper insight

RESULTS

- Error for male is consistently less than that for female distribution
- Addition of the audio modality improves performance, but has the highest amount of bias
- 13 out of 26 features have a statistically significant difference for the male and female distributions under the two-tailed T-test ($\alpha = 0.05$) after the Bonferroni correction

	$\Delta G = MSE_F - MSE_M$	t
$\tau = 3$	0.38	0.16
$\tau = 7$	0.26	0.18
$\tau = 35$	0.18	
$\tau = 30$		
MDRM(A)	0.38	
MDRM(T)	0.38	0.12
MDRM(M)	0.30	0.11
	0.28	0.16
		0.14

Analysis Features	P value	Statement
Pitch Analysis		
Mean Fundamental Frequency (F0)	<0.001	*
SD Fundamental Frequency (F0)	<0.001	*
Number of pitch	<0.001	*
Number of words	<0.001	*
Degree of word words	<0.001	*
Minimum F0	<0.001	*
Maximum F0	<0.001	*
Mean F0	<0.001	*
Vowel to Length Ratio	<0.001	*
Vowel to Total Ratio	<0.001	*
Amplitude Analysis		
Mean Intensity	<0.001	*
SD Intensity	<0.001	*
Maximum Intensity	<0.001	*
Minimum Intensity	<0.001	*
Rate Analysis		
Local Average Time	<0.001	*
Relative Average Formulation Time	<0.001	*
Point Formulation Quantum Time	<0.001	*
gap time	<0.001	*
Local Silence	<0.001	*
Local Silence	<0.001	*
avg3 Silence	<0.001	*
avg5 Silence	<0.001	*
avg7 Silence	<0.001	*
avg9 Silence	<0.001	*
Demographic Analysis		
Demographic Noise Ratio	<0.001	*

Beyond fair pay: ethical implications of NLP crowdsourcing

**Beyond Fair Pay:
Ethical Implications of
NLP Crowdsourcing**

Boar Shmueli^{1,2}, Jan Fell³, Soumya Ray³, Lun-Wei Ku¹
Academic Editor: Northeastern University
NAACL 2022

Crowdsourcing in NLP research

Year	Accepted Papers				Papers Using Crowdsourcing	Payment Mentioned	IRB Mentioned
	ACL	EMNLP	NAACL	All			
2015	219	212	186	810	36 (45%)	0	0
2016	264	183	160	597	31 (52%)	15 (25%)	0
2017	302	333	—	625	37 (6%)	12 (21%)	2
2018	281	345	332	1282	125 (11%)	11 (3%)	1
2019	602	681	423	1798	183 (11%)	32 (17%)	5
2020	773	754	—	1933	182 (12%)	42 (23%)	5
Total	2768	2885	1123	6776	763 (10%)	122 (17%)	14

Are Crowdworkers **Human Subjects**?

- Do researchers obtain information about the worker?
- Do researchers obtain identifiable private information (IPI)?

**Are IRBs
universal? ?**

**Are worker IDs
Identifiable
Private
Information**

Risks and Harms for Crowdworkers

- Inducing psychological harm
- Exposing sensitive information of workers
- Unwittingly including vulnerable populations
- Breaching anonymity and privacy
- Triggering addictive behaviour

1/5
Inducing psychological
harms

2/5
Exposing sensitive
information of workers

3/5
Unwittingly including
vulnerable populations

4/5
Breaching anonymity and
privacy

5/5
Triggering addictive
behaviour

Case study: deontological ethics in NLP



Overview

- Ethical issues in NLP should be analyzed within the ethical frameworks which have been studied extensively in philosophy.
- Goal:** To show NLP practitioners how philosophical theories of ethics can be directly applicable to NLP.
- Which tasks have important ethical implications?
- What factors and methods are preferable in ethically solving this problem?

Our primary contributions are:

- Providing an overview of two deontological principles along with a discussion on their limitations with a special focus on NLP.
- Illustrating four specific case studies of NLP systems which have ethical implications under these principles and providing a direction to alleviate these issues.

Some of the problems and suggestions we make in this paper are already known to the community, yet the aim of this paper is to identify particular problems as specifically ethical issues rather than simply technical or practical issues.

Deontological Ethics

Deontological ethics is a family of ethical theories which holds that ethical action is determined by rules and rights. This is contrast to ethical theories like consequentialism (e.g., utilitarianism) which is based on outcomes of actions. We select deontological ethics for this paper because

- It is a widely studied category of ethics.
- Rules and rights provide a systematic basis for NLP practitioners to work from.
- Rights and duties which apply to everyone equally fit well with the widely used legal concept of rule of law.

We have selected the generalization principle and informed consent as the two principles for this case study as the former is abstract and far-reaching while the latter is more concrete and focused.

References

- [1] Bassukil, A. et al. "Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-Shot Commonsense Question Answering". In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), 2021.
- [2] Selenia Drago et al. "Hey Google Is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction". In: Proceedings of the 2017 Conference on Interaction Design and Children, 2017.
- [3] Johnson et al. "Kant's Moral Philosophy". In: The Stanford Encyclopedia of Philosophy, 2019.
- [4] J. Pugh. Autonomy, Rationality, and Contemporary Bioethics [Internet]. Oxford University Press, 2020.
- [5] Maarten Sap et al. "Social Bias Profiles: Reasoning about Social and Power Implications of Language". In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

Principle 1: Generalization

The generalization principle originated as a formulation of Immanuel Kant's categorical imperative, the central rule of his ethical theory. It is stated as [3]:

An action A taken for reasons R is unethical if and only if a world where all people perform A for reasons R logically contradicts R .

Example

A : breaking a contract

R_1 : they believe the other party will uphold the contract

R_2 : will gain an advantage by breaking the contract

If everyone were to break contracts (A) for these reasons, no one would enter into a contract believing the other party would uphold the terms, thereby contradicting R_1 .

Case Study 1: Question Answering

A : a QA system provides responses to users' questions based on heuristics without the ability to justify or explain its answer (A)

R_1 : the user does not know the answer to the question

R_2 : the user will trust the response of the QA system.

A is unethical because if all QA systems were unable to give explanations to their answers, especially incorrect answers, users would lose trust in the systems, contradicting R_2 .

Case Study 2: Content Moderation

A : a content moderation system for social media flags certain content as objectionable based on superficial features

R_1 : the objectionable is identifiable by the system

R_2 : deploying the system reduces such offensive content

A is unethical because if all content moderation systems used surface-level features alone, authors of offensive content could simply express the same meaning in a different way and avoid detection, contradicting R_1 .

The way forward QA methods which generate answers from explainable representations such as knowledge graphs can accurately display reasoning to the user [1]. Thus the user can simply see an error reasoning as the cause for the incorrect answer.

Principle 2: Informed Consent

Informed consent is a special case of respect for autonomy which holds that person generally has the right to decide what they do and what happens to them. We use the following formulation [4]:

- A Person A potentially performs some act X on person B which would normally infringe on B 's autonomy.
- It is unethical for A to perform X unless:

- B is sufficiently informed as to the nature of X and its consequences.
- On the basis of this information, B themselves make the decision to permit A to perform X .

Example A person (B) has a right to decline avoidable harm to their body. A doctor (A) may propose to perform an experimental treatment (X) on B which has both risks and potential benefits. It is unethical for A to perform X unless B both understands the risks of X and consents to X .

Case Study 3: Machine Translation

A : Person (B) has the right to speak for themselves.

- If B and another person do not share a language, B may use an MT system (A) which speaks on behalf of B (X).

X is unethical unless:

- B understands the failure modes of the MT system and what type of misunderstandings might occur. E.g., translating an idiom literally may convey the wrong meaning.
- B either requires the MT system not to give mistranslations or acknowledges that these translation failures and their consequences are acceptable.

The way forward In order to obtain a proper informed consent, the user of potential misunderstandings, the MT system must both aware of quality of its output and of the cultural contexts of the input and output language.

Case Study 4: Dialogue Agents

A : Parents (B) have a right to restrict whom their (young) children speak with and what they talk about.

- A smart assistant (A) (e.g., Amazon Alexa) installed in family's house may speak with children in the household (X).

X is unethical unless:

- The parents (B) understand the how their children might interact with the smart assistant. For example, parents may not be aware that young child would see a smart assistant as being capable of feelings or as a trustworthy source of answers [2].
- The parents (B) must be able to control what type of interactions the smart agent is allowed to have with their children (X).

The way forward Regarding (1), the smart assistants (or their developers) must provide information to the parents regarding the ways in which the smart assistant might interact with children and what the effects might be. Regarding (2), parents must be able to limit what sort of interactions the smart assistant has with their children. In order to do this, the smart assistant must be aware when a child is talking to it.

- Goal: leverage a large body of work on ethics. See how we can apply them to NLP.
- Deontological framework for NLP
 - Generalization principle (categorical imperative: An action A is ethical iff a world where all people performing A is conceivable)
 - Respect for Autonomy
- Reasonable, clear ethical rules, "rule of law"
- Four case studies: QA, MT, detecting objectionable content, dialogue systems
 - Which tasks have important ethical implications?
 - What factors and methods are preferable in ethically solving this problem?

On transferability of bias mitigation effects in language model fine-tuning



On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning

Xisen Jin¹, Francesco Barbieri², Brendan Kennedy¹, Aida Mostafazadeh Davani¹, Leonardo Neves², Xiang Ren¹
University of Southern California¹, Snap Inc.²



Problem statement

Can we debias an upstream model **once** and **retain the effect of bias mitigation** when fine-tuned in later **downstream applications**?

Bias of Interest	Bias Factors	Advantages
Disparate model performance between different groups 	Group identifier bias / AAVE dialect bias / Gender bias	<ul style="list-style-type: none"> Significant reduction of efforts in downstream Encourages broader application of bias mitigation

Results and Analysis

Tasks: Hate speech (GHC, Stf) / Toxicity (FDCL, DWMW) / Occupation Prediction (Biasbios) / Coreference (OntoNotes 5.0)

Compared methods: Vanilla, Downstream bias mitigation, Vanilla Transfer Learning, UBM

Same Domain & Task	Cross Domain & Task		Multiple bias factors
	GHC Part A → GHC Part B	GHC → Stf & FDCL → DWMW	

- UBM notably reduces bias compared to Vanilla / Vanilla Transfer Learning
- Does not rival direct downstream bias mitigation
- It is possible to reduce multiple bias factors via UBM across domain and tasks
- These effects are not automatic for each new dataset added

Upstream Bias Mitigation (UBM) Framework

Setups

Task & Domain Similarity	Number of Bias Factors Mitigated
<ul style="list-style-type: none"> Same domain & task Cross domain & task 	<ul style="list-style-type: none"> One bias factor Multiple bias factor

Conclusions

We show that the **effects of bias mitigation** are indeed **transferable** in fine-tuning language models.

- Though UBM does not rival directly mitigating bias on the downstream task, it is more efficient and accessible.

Future works can study algorithms to improve UBM and mitigate bias in a more reliable way.

Privacy regularization: joint privacy-utility optimization in LMs

1. Problem

Neural language models are known to have a high capacity for memorization of training samples. This may have serious privacy implications when training models on user content such as email correspondence.

Unintended Memorization of Secrets
My credit card number is 4403 2212 8563 2345

Proposed solution

We propose two privacy regularization methods, based on adversarial training and a novel privacy loss term, to jointly optimize for privacy and utility of recurrent language models. The main idea of our regularizers is to prevent the last hidden state representation of the language model for an input sequence from being linked back to the sensitive attribute we are trying to protect.

2. Motivation

We show that differential privacy can have shortcomings in addressing this problem, for the reasons below:

- DP is not context-sensitive: Cannot explicitly define protected attribute and wire it in the loss
- DP is not suitable for correlated/repeated data
- DP has disparate impact
- DP training is 10-15X slower, and much more cumbersome to tune

3. Results

4. Results

Our results show that our regularization can be as effective as differential privacy, and more effective in some special cases. We also show that our regularizers do not have the disparate impacts of differential privacy, on utility.

Papers in semantics

Session 1E (sentence-level, textual inference)

Unifying cross-lingual SRL with heterogeneous linguistic resources

The poster is titled "Unifying Cross-Lingual Semantic Role Labeling with Heterogeneous Linguistic Resources" by Simone Conia, Andrea Baciu, and Roberto Navigli from the Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome. It features a dark header with the title and authors, and a dark footer with logos for SAPIENZA NLP, the European Union, and the project's funding.

Semantic Role Labeling (SRL)

A brief introduction: Semantic Role Labeling is the task of automatically addressing questions like "Who did fiber to fibers, Where, When, Why and How?" (Elman et al., 2007).

Model Overview

Learning from heterogeneous linguistic resources: Recent evaluations have shown that our approach can outperform state-of-the-art systems across multiple languages.

Model Architecture: universal encoders

In the universal encoder, predicate sense and predicate argument representations are shared across languages.

Experiments on Dependency-based SRL (CoNLL-2009)

How does our approach fare? Comparison with He et al., 2019

Cross-Lingual SRL in low-data settings: Benefits of cross-lingual approach are evident in low-data settings.

1-Shot Cross-Lingual SRL: 1-shot learning = 1 sentence / predicate sense

Analysis, Discussion and Takeaways

One model for many linguistic resources: Multiple arguments in a single forward pass. A model needs to automatically compare different inventories and linguistic theories.

Aligning meaning across inventories: What our model implicitly learns.

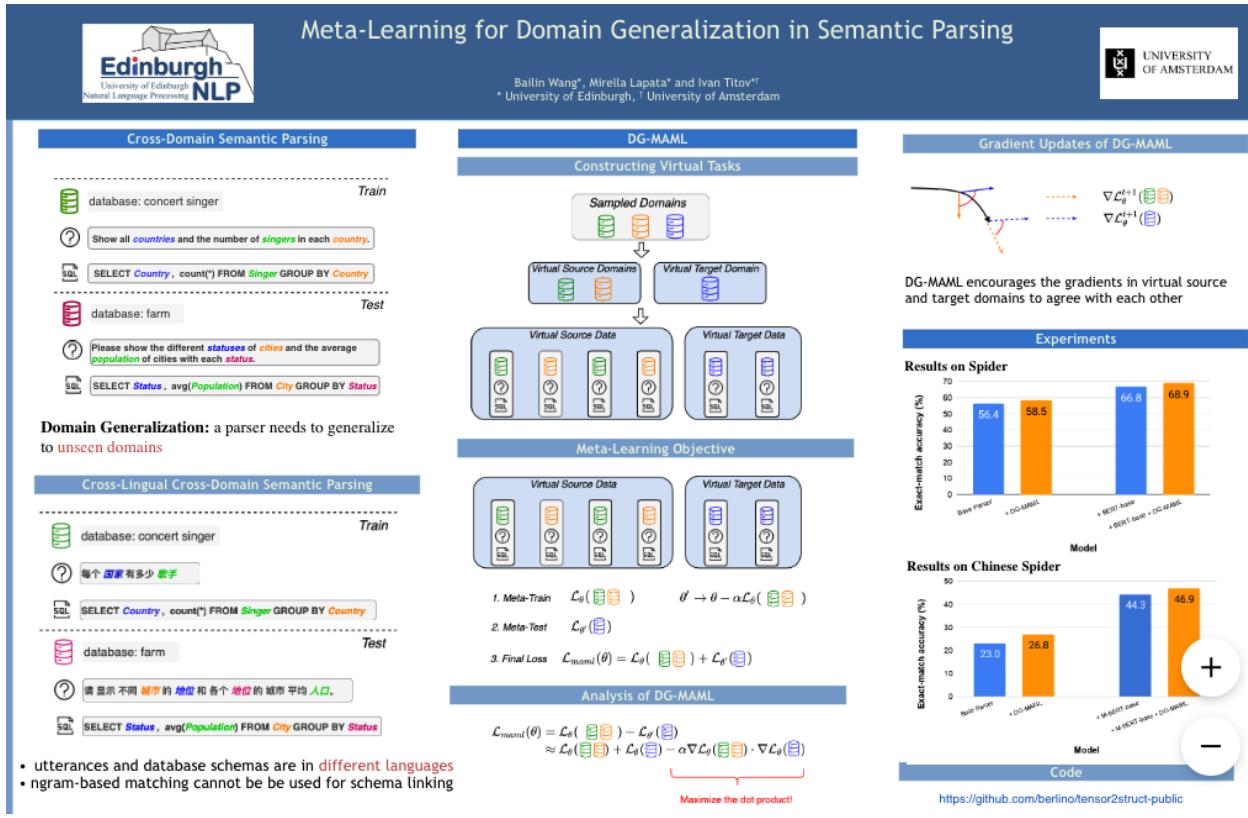
Conclusion

Key takeaways:

- One inventory (SL)
- aligning diverse resources
- NAACL reproducible code

NAACL 2021

Meta-learning for domain generalization in semantic parsing



Session 4C (sentence-level, textual inference)

Understanding by understanding not: modeling negation in LMs



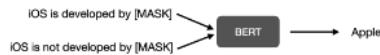
Understanding by Understanding Not: Modeling Negation in Language Models

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni and Aaron Courville
Mila, Université de Montréal

Abstract

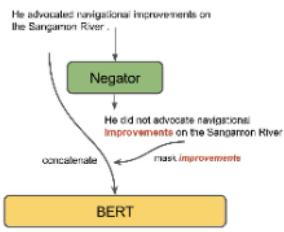
Negation is a core construction in natural language. Despite being very successful on many tasks, state-of-the-art pre-trained language models often handle negation incorrectly. To improve language models in this regard, we propose to augment the language modeling objective with an unlkelihood objective that is based on negated generic sentences from a raw text corpus. By training BERT with the resulting combined objective we reduce the mean top 1 error rate to 4% on the negated LAMA dataset. We also see some improvements on the negated NLI benchmarks.

Motivation



- Negation plays a key part in several language understanding tasks, e.g. sentiment analysis, question answering, NLP and knowledge base completion
- Pre-trained language models achieve SOTA on various tasks
- However, Kassner and Schütze (2019) show that PLMs, such as BERT, cannot correctly distinguish between the negated and non-negated fill-in-the-blank queries

Unlikelihood with reference



$$L_{UL} = -\log(1 - p(\text{improvements} | X_{1:T}))$$

- we contextualize each sentence by concatenation
- negation is always false in the "small world" created by its context

Syntactic Negation Augmentation

Many fonts then made the right leg vertical

Pattern: {S ; cpos : /, *Tense=Past, */=A >/ nsubj | csubj=/E
[]-subject?> obj {tag : /NN, */=object}

Actions:

Type: Insert

Token: did

Type: Insert

Token: not

Type: Lemmatize

Many fonts then did not make the right leg vertical

Experimental Results

Model	SQuAD	ConceptNet	T-REx	Google-RE
BERT	13.53	15.85	29.10	10.24
BERT + XL	13.64	15.84	29.28	10.27
BERTNOT	13.87	15.49	29.25	10.31

mean precision at 1 ($p@1$) for LAMA queries (lower is better)

Query	Top 3 words from BERT	Top 3 words from BERTNOT
iOS is developed by [MASK]	Apple, Google, Microsoft	Apple, Google, Microsoft
iOS is not developed by [MASK]	Apple , Google, Microsoft	Apple, Google, Microsoft
The majority of the amazon forest is in [MASK]	Brazil, Bolivia, Madagascar	Brazil, Bolivia, Mexico
The majority of the amazon forest is not in [MASK]	cultivation, Brazil , Mexico	cultivation, Mexico, France
Charles Nodier died in [MASK]	Paris, Rome, office	Paris, Rome, France
Charles Nodier did not die in [MASK]	Paris , office, France	van, error, doubt
Mac OS is developed by [MASK]	Apple, Microsoft, Intel	Apple, Microsoft, Israel
Mac OS is not developed by [MASK]	Apple , Microsoft, IBM	Apple, Microsoft, IBM

Model	RTE	SQuAD
BERT	76.04_{±0.05} (69.68 _{±1.35})	65.47 _{±0.15} 74.47_{±0.29}
BERTNOT	89.00 _{±0.18}	89.10 _{±0.18} 45.96_{±0.01} 84.31_{±0.29} 60.89_{±0.01}

Assesses on original dev sets and new splits containing negation from Hosseini et al. (2020) (Refer)

Disentangling semantic and syntax in sentence embeddings with pre-trained LMs



Disentangling Semantics and Syntax in Sentence Embeddings with Pre-trained Language Models

James Y. Huang, Kuan-Hao Huang, Kai-Wei Chang
University of California, Los Angeles

Motivation

- Semantic sentence embedding** models map sentences with closer semantics into closer embedding vectors.
- Sentence embeddings from pre-trained language models encode **rich but entangled semantic and syntactic information**.
- Goal:** improve semantic sentence embeddings from pre-trained language models by **learning to disentangle semantics and syntax**.

Disentangling Semantics and Syntax

How to learn the distinction between semantics and syntax?

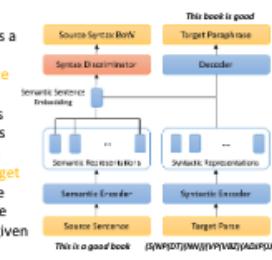
- Paraphrase pairs always have close semantics but often come in different syntax.
 - We propose to train a semantic sentence embedding model as part of a **syntax-guided paraphrasing model**.
 - Syntactic guidance encourages the semantic nature of sentence embeddings
-

ParaBART: Learning Disentanglement from Paraphrases

ParaBART improves semantic sentence embeddings from pre-trained BART by learning **semantics-syntax disentanglement** from **paraphrase pairs**.

Paraphrasing model

- Semantic encoder** learns a semantic sentence embedding from a **source sentence**.
- Syntactic encoder** learns syntactic representations from a **target parse tree**.
- Decoder** generates a **target paraphrase** of the source sentence that follows the syntax specified by the given parse tree.



Syntax Discriminator

- Attempts to **recover source syntax** from the semantic embedding by minimizing an adversarial syntax prediction loss
- Exposes syntactic information encoded in semantic sentence embeddings

Training Objective

- Syntax discriminator:** predict source syntax from semantic sentence embedding
- Paraphrasing model:** generate target paraphrase + "fool" the syntax discriminator

$$\min_{E_{\text{sem}}, E_{\text{syn}}, D_{\text{adv}}} \left(\max_{D_{\text{adv}}} (L_{\text{para}} - \lambda_{\text{adv}} L_{\text{adv}}) \right)$$

Experimental Results

Unsupervised Semantic Textual Similarity (STS)

- Goal:** estimate semantic similarity of two sentences by computing the cosine similarity of their sentence embeddings.
- Strong performance across STS tasks
- Significant improvement from pre-trained BART embeddings

Model	STS 12	STS13	RF508	STS16	STS14	STS-B	Avg.
Aug. BERT embeddings	46.9	52.8	57.2	63.6	68.5	53.8	55.5
Aug. BART embeddings	50.8	52.8	56.1	63.0	68.5	53.8	54.2
InferSent	43.0	58.4	70.0	71.6	73.1	60.7	60.7
Universal Sentence Encoder	41.9	54.3	67.2	73.8	74.3	66.0	66.0
Sentence-BERT	41.8	63.5	73.6	74.8	73.9	74.2	69.7
BERT	44.6	63.5	73.2	74.8	70.1	74.1	70.6
ParaBART	50.9	52.2	58.9	79.6	79.1	75.1	75.4
- w/o Adversarial Loss	47.5	70.0	75.8	80.9	80.0	78.7	76.5
- w/o Adversarial Loss & Syntactic Guidance	46.4	68.3	72.6	80.0	78.6	75.4	73.2

Syntactic Probing

- Goal:** investigate to what degree our semantic sentence embeddings can be used to predict syntactic properties.
- Lower accuracy on these tasks suggests **less syntax being encoded** in semantic sentence embeddings.
- ParaBART significantly reduces the amount of syntactic information in semantic sentence embeddings.



Temporal reasoning on implicit events from distant supervision

Temporal Reasoning on Implicit Events from Distant Supervision



Ben Zhou^{1,2}, Kyle Richardson², Qiang Ning³, Tushar Khot², Ashish Sabharwal², Dan Roth¹



AI2

¹University of Pennsylvania, ²Allen Institute for AI, ³Amazon

1. Contribution

TRACIE (TempoRAI Closure Inference)

- A temporal relation benchmark on **implicit events**
- 5.5k entailment instances
- Test both start and end time of events
- Improved Models on TRACIE
- PatternTime:** Trained from pattern-based distant signals collected from unannotated free-texts. These signals are designed for implicit events.
- SymTime:** Neural-symbolic model that symbolizes Allen's interval algebra. Infers end time with start time and duration estimations.

2. Motivation

Systems should be able to construct latent timelines

- With both explicit and **implicit events**
- To show real understanding of situations

Yet, no previous work focus on this problem and propose benchmarks, analysis or systems.

3. TRACIE Construction

Stage 1: Implicit event generation

- Sample context stories from RDCStories.
- Annotators write implicit events according to relatedness requirements.
- Stage 2: Hypothesis generation
- Collect a pool of explicit events from annotator paraphrases and SRL-based extractions.
- Randomly pair implicit-explicit event pairs with comparator and query.
- Stage 3: Instance Labelling
- Annotators always compare with the implicit event's start/end time with the explicit event's start time. This produces maximum accuracy, as explicit event's start time is easy to ground.
- 4 different annotators label each instance, their majority agreement (dropped if non-exist) is used as the final label.
- Expert: 94% agreement, 98% resolved accuracy

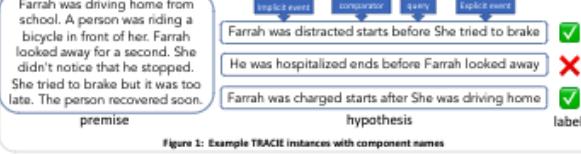


Figure 1: Example TRACIE instances with component names

3. PatternTime

We further pre-train a T5-large model with two distant supervision sources from free texts.



Within-sentence extraction relies on SRL model and direct mentions of before/after

Cross-sentence extraction relies on temporal expression (dates, hours) mentions, improves implicit event understanding (distant-independent) and produces relative interval estimations.

Input: two event phrases, **Output:**

- 1) A binary label for start temporal relations;
- 2) A probability vector indicating which duration unit is closest to the relative interval.

4. SymTime

3.5. SymTime

Overview. SymTime infers end time as start time + duration with an end2end neural-symbolic model.

Assume the first event's start time start_1 , duration duration_1 , and the second event's start time start_2 , we want to compute $\text{sign}(\text{duration}_1 - (\text{start}_2 - \text{start}_1))$

Sub-modules.

- PatternTime provides $\text{start}_1 - \text{start}_2$.
- For duration_1 , we train a duration model with the distant supervision collected from previous work, which predicts a probability vector over the same set of duration units.
- Computation.** Both PatternTime and the duration module produce probability vectors. To get a single number, we dot product them with a constant incremental vector c to get a weighted mean $\langle c | x \rangle$. We use the binary label from PatternTime and apply a tanh function to get a sign close to either -1 or 1 from probabilities $\langle g | x \rangle$. Values are computed to resemble the formula above.
- Zero-shot Version.** As both modules are pre-trained with distant signals, SymTime can be applied without task-specific supervision.
- We call this version **SymTime-ZS**



Figure 3: SymTime computation for end time queries

5. Experiments

TRACIE experiments

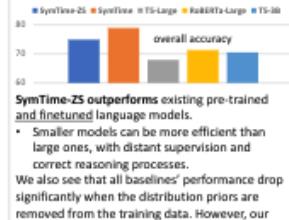
- Split of 20/80 train/test ratio.
- Remove the priors bias over comparator-query-label distributions.
- T5-large is our base model and main target of comparison



PatternTime improves much on start-time comparisons, because of the distant supervision collected automatically via patterns.

SymTime improves on end time comparisons with its symbolic computation.

SymTime outperforms the larger **TS-3B**, showing the benefit of distant supervision + reasoning.



(more experiments in the paper)

Session 5E (stylistic analysis)

Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa

DOES SYNTAX MATTER?

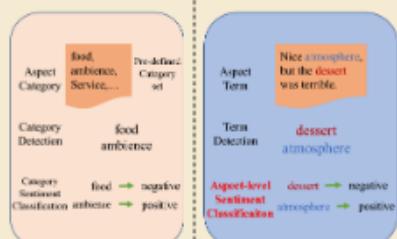
A strong baseline for
Aspect-based Sentiment
Analysis with RoBERTa



Junqi Dai*, Hang Yan*, Tianxiang Sun, Pengfei Liu, Xipeng Qiu

INTRODUCTION

The atmosphere is nice, but the dessert was dreadful.

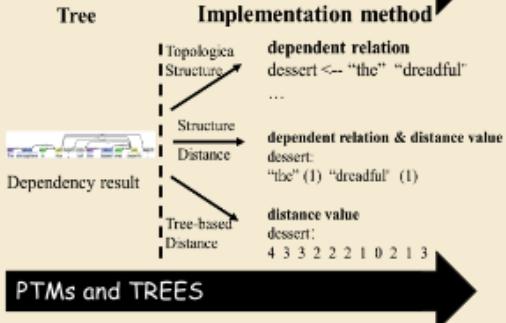


The atmosphere is nice, but the dessert was dreadful.

Aspect Category	Aspect Term
food, ambience, Service, ...	Nice atmosphere, but the dessert was terrible.
Aspect Term	Aspect Category
dessert atmosphere	food ambience

Aspect-level sentiment classification (ALSC) aims to do the fine-grained sentiment analysis towards. Specifically, for one or more aspects in a sentence, the task calls for detecting the sentiment polarities for all aspects.

AISC MODEL and TREES

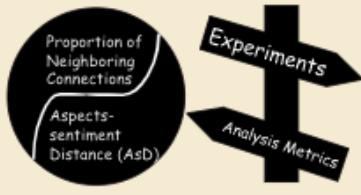


Questions:

- Tree induced from PTMs vs. Tree from dependency parser ?



- Tree induced from PTMs vs. Tree from task fine-tuned PTMs ?



ALSC models

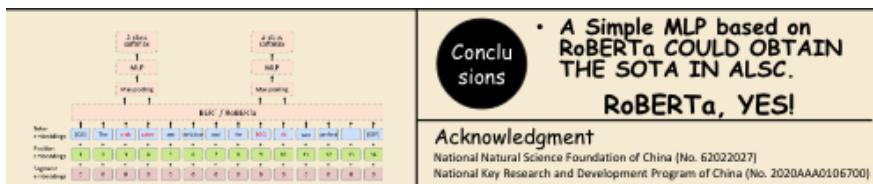
- ASGCN
 - Topological Structure
- PWGN
 - Relative Distance
- RGAT
 - Structure & Distance

Main Experiments

Incorporate all trees with all ALSC models.

Trees

- Dep. (Dependency tree)
- Left/Right chain
- BERT Induced Tree
- ROBERTa Induced Tree
- FT-BERT Induced Tree
- FT-ReBERTa Induced Tree



Domain divergences: a survey and empirical analysis

Domain Divergences: a Survey and Empirical Analysis

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann
School of Computing, National University of Singapore

Introduction

- Domain Divergence is a primary tool in measuring domain shift
- In this work we come up with a Taxonomy of divergence measures: Geometric, Information Theoretic, Higher Order
- We identify 3 use cases of divergences. a) Data Selection b) Learning Domain invariant representations c) Decision in the wild
- One major use case is to predict the drop in performance.
- Which measure best predicts the drop? Till now, word distributions and word embeddings are used. Is there any advantage of using contextualised word representations for predicting drops?

Methodology

Fine tune DistilBERT model on source domain \mathcal{S}

Performance Drop = Accuracy on Test data of \mathcal{S} - Accuracy on Test data of \mathcal{T}

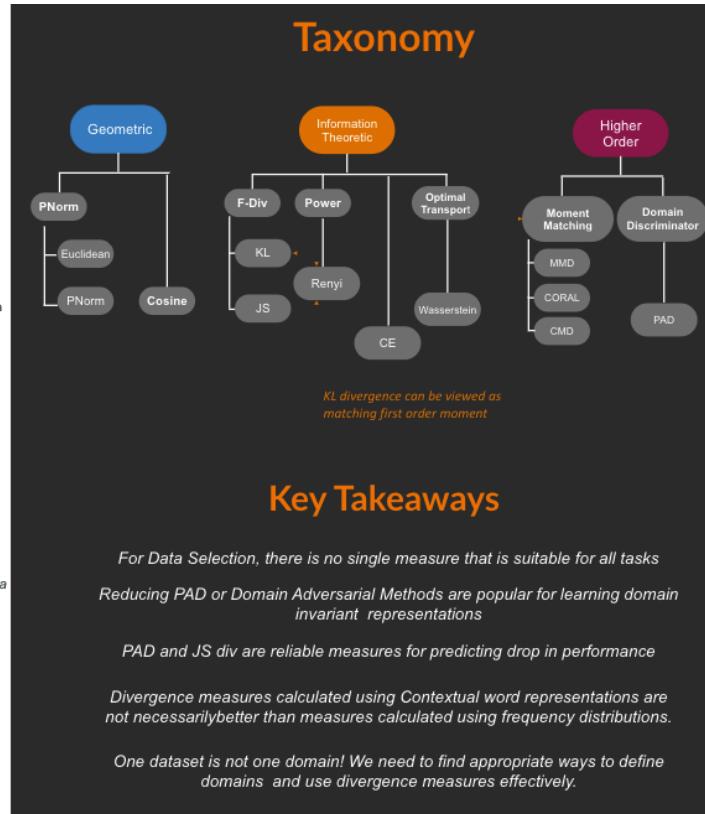
Correlation of performance drop with the divergence between domains

Datasets

- POS- 5 corpora from English Web Tree Bank Corpus, NER-8 corpora, Sentiment Analysis: Amazon Review Dataset with 5 categories

Divergence Measures

- 12 divergence measures used in the literature



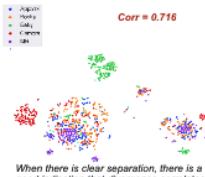
Divergence/Task	POS	NER	SA
Cos	0.016	0.223	-0.012
KL-Div	0.394	0.384	0.715
Js-Div	0.407	0.464	0.709
Renyi Div	0.392	0.382	0.716
PAD	0.477	0.426	0.538
Wasserstein	0.378	0.463	0.448
MMD-RQ	0.248	0.495	0.614
MMD-Gaussian	0.402	0.221	0.543
MMD-Energy	0.244	0.447	0.521
MMD-Laplacian	0.389	0.273	0.623
CORAL	0.349	0.484	0.267

Correlation between divergence Measure and performance drop

Divergence/Tasks	POS	NER	SA
Cos	-1.76×10^{-1}	-2.49×10^{-1}	-2.01×10^{-1}
KL-Div	-	-	-
Js-Div	-4.5×10^{-1}	-8.4×10^{-1}	2.04×10^{-1}
Renyi Div	-	-	-
PAD	-	-	-
Wasserstein	-2.11×10^{-1}	-2.36×10^{-1}	-1.70×10^{-1}
MMD-RQ	-4.11×10^{-1}	-3.04×10^{-1}	-1.70×10^{-1}
MMD-Gaussian	4.25×10^{-1}	2.57×10^{-1}	-8.45×10^{-1}
MMD-Energy	-9.84×10^{-1}	-1.14×10^{-1}	-8.48×10^{-1}
MMD-Laplacian	-1.67×10^{-1}	4.26×10^{-1}	-1.08×10^{-1}
CORAL	-2.34×10^{-1}	2.78×10^{-1}	-1.41×10^{-1}

Silhouette coefficients with different divergence measures.

Dataset is not a domain



When there is clear separation, there is a good indication that divergence correlates well with performance drop

Session 8C (sentence-level, textual inference)

Learning from executions for semantic parsing



Learning from Execution for Semantic Parsing

Bailin Wang*, Mirella Lapata* and Ivan Titov†
 * University of Edinburgh, † University of Amsterdam

Semantic Parsing

Data Example

Domain: Restaurant
NL: list all 3 star rated thai restaurants
Program: SELECT restaurant WHERE star.rating = 3 AND cuisine = thai

Task: mapping a natural language (NL) utterance to its corresponding executable program

Motivation for Semi-Supervised Learning

Example

NL: list all 3 star rated thai restaurants

Candidate Programs	Gold	Exe
SELECT restaurant WHERE star.rating = thai	X	X
SELECT restaurant WHERE cuisine > 3	X	X
SELECT restaurant WHERE star.rating = 3	X	✓
SELECT restaurant WHERE star.rating = 3 AND cuisine = thai	✓	✓

- Not all candidate programs make sense.
- Executability is a weak yet free learning signal.

Maximum Marginal Likelihood

$$\mathcal{L}_\theta(x) = -\log \sum_y R(y)p(y|x, \theta)$$

where x, y denote NL and program respectively. $R(y)$ returns 1 if y is executable; it returns 0 otherwise.

Divided Program Space

	Seen Programs	Unseen Programs
Executable Programs	P_{SE}	P_{UE}
Non-Executable Programs	P_{NE}	P_{UN}

Beam search can help us see a subset of programs

Posterior Regularization

We assume a constrained family of distribution \mathcal{Q} : for any $q \in \mathcal{Q}$, $E_{q(y|x)}[R(y)] = 1$.

For a semantic parser $p(y|x, \theta)$, the objective of posterior regularization (Ganchev et al., 2010) is to penalize the KL divergence between \mathcal{Q} and p .

New Objectives

$$q_{repulsion}^{t+1}(y) = \begin{cases} \frac{p(y|x, \theta^t)}{1-p(P_{SE})} & y \notin P_{SE} \\ 0 & \text{otherwise} \end{cases}$$

$$q_{attract}^{t+1}(y) = \begin{cases} \frac{p(P_{SE|y})}{p(P_{SE})} p(y|x, \theta^t) & y \in P_{SE} \\ \frac{p(y|x, \theta^t)}{p(P_{UN})} & y \in P_{UE} \cup P_{UN} \\ 0 & y \in P_{NE} \end{cases}$$

$$q_{prime}^{t+1} = \text{SparseMax}_{y \in P_{SE}} (\log p(y|x, \theta^t))$$

EM Algorithm for Optimizing PR

E-step

find the q that is closest to p

M-step

optimize p wrt. the parameters

Optimizing PR is equivalent to optimizing MML

Results on Overnight

Analysis: Length Ratio

<https://github.com/berfinol/tensor2struct-public>

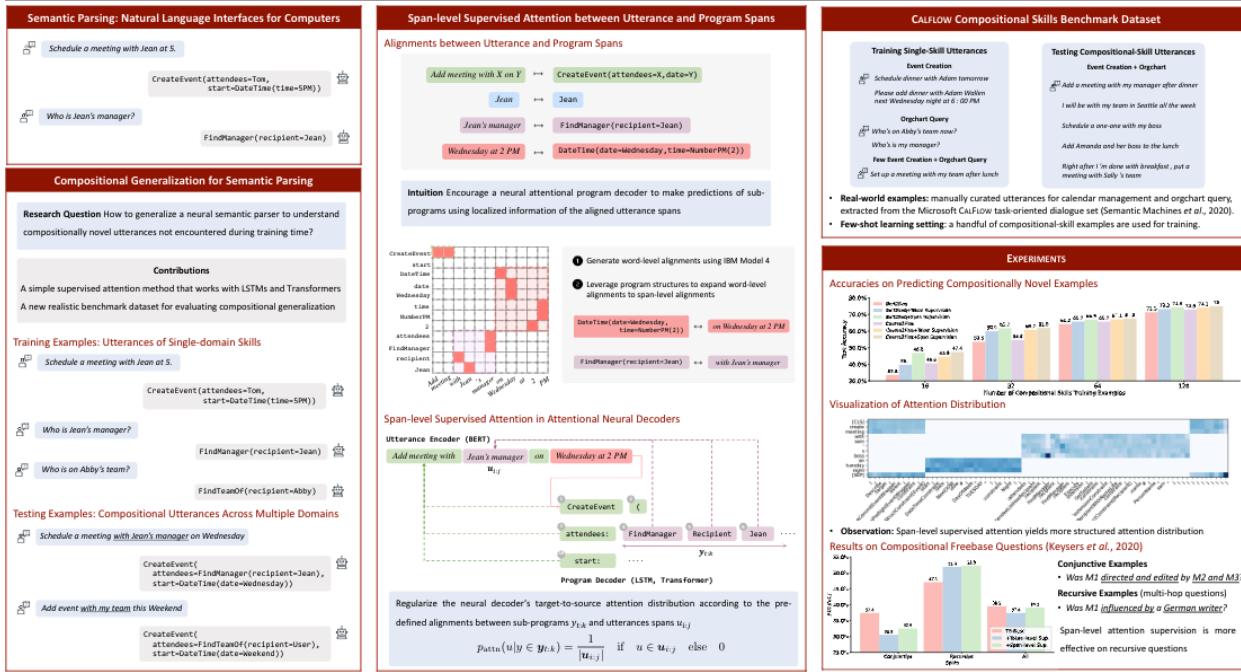
Compositional generalization for neural semantic parsing via span-level supervised attention



COMPOSITIONAL GENERALIZATION FOR NEURAL SEMANTIC PARSING VIA SPAN-LEVEL SUPERVISED ATTENTION

Pengcheng Yin¹, Hao Fang², Graham Neubig², Adam Paulin², Antonios Platanios², Yu Su², Sam Thomson², Jacob Andreas²

¹Language Technologies Institute, Carnegie Mellon University ²Microsoft Semantic Machines



Incorporating external knowledge to enhance tabular reasoning

Incorporating External Knowledge to Enhance Tabular Reasoning

J. Neeraja⁽¹⁾, Vivek Gupta⁽²⁾, Vivek Sri Kumar⁽²⁾

(1) IIT Guwahati; (2) University of Utah



1. Tabular Inference Problem

- Inference task where premises are tabular in nature
- Given a premise table determine hypothesis is true (entailment), false (contradiction), or undetermined (neutral), i.e. tabular natural language inference.

New York Stock Exchange	
Type	Stock exchange
Location	New York City, New York, U.S.
Founded	May 17, 1792; 226 years ago
Currency	United States dollar
No. of listings	2,400
Volume	US\$20.161 trillion (2011)

H1: NYSE has fewer than 3,000 stocks listed.
 H2: Over 2,500 stocks are listed in the NYSE.
 H3: S&P 500 stock trading volume is over \$10 trillion.

- Example InfoTabS dataset (Gupta et al., 2020), H1: entailed ; H2: contradictory ; H3: neutral

2. Motivation

- Recent work mostly focuses on building sophisticated neural models.
- How will models designed for the raw text adapt for tabular data?
- How to represent data and incorporate knowledge into these models?
- Can better pre-processing of tabular information enhance table comprehension?

3. Challenges

- Poor Table Representation
- Missing Lexical Knowledge
- Presence of Distracting Information
- Missing Domain Knowledge

Main Question

Can we fix the above problems by changing how tabular information is provided to a standard model?

4. Poor Table Representation

- Using universal template → Most sentences are ungrammatical or non-sensible
- X** The Founded of New York Stock Exchange are May 17, 1792; 226 years ago.

Better Paragraph Representation

- Entity specific templates : use value entity types DATE, MONEY or CARDINAL or BOOL

✓ New York Stock Exchange was founded on May 17, 1792; 226 years ago.

- Add category information.

New York Stock Exchange is an organization

More grammatical and meaningful sentences

5. Missing Lexical Knowledge

- Limited training data → affects interpretation of hypernym words such as *fewer*, *over* and negations.

Implicit Knowledge Addition

Can pre-training on large NLI dataset help?

- Pre-training with MNLI data
- Then, fine-tune on InfoTabS

Espresso model to diverse lexical constructions. Representation is better tuned for the NLI task.

6. Distracting Information Issue

- Only select rows are relevant for a given hypothesis. E.g. **No. of listings** is enough for H1 and H2.
- Due to BERT tokenization limit, useful rows in the longer tables dropped.

Distracting Row Removal

- Select only rows relevant to hypothesis.
- Use Alignment based retrieval algorithm with fastText vectors (Yadav et al. (2019, 2020))

E.g. for H1, H2, new prune table :

New York Stock Exchange	
No. of listings	2,400

7. Missing Domain Knowledge

- For H3, we need to interpret **Volume** in financial context.

✓ In capital markets, volume, is the total number of a security that was traded during a given period of time.

rather than

X In thermodynamics, volume of a system is an extensive parameter for describing its phase state.

Explicit Knowledge Addition

- Add explicit information to enrich keys.
- This improves model's ability to disambiguate meaning of keys.

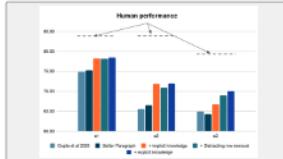
Approach

- Use BERT on wordnet examples to find key embeddings
- Get key embeddings from premise using BERT
- Find the best match and add it definition to premise.

Add to the table in the end for H3

Volume: total number of a security that was traded during a given period of time.

8. Experimental Results



- Significant improvement in adversarial α_2 and α_3 dataset

Ablation Study: All changes are needed, knowledge addition being the most important.

9. Conclusion

- Proposed pre-processing lead to significant improvements

Propose approach beneficial for adversarial α_1 and α_2 dataset

Solutions applicable to question answering and generation problems with both the tabular and textual inputs

Proposed modifications should be standardized across other table reasoning tasks

Data and Software:

<https://infotabs.github.io>

10. References

- Gupta et. al. INFOTABS: Inference on Tables as Semi-structured Data. ACL'20.
- Yadav et. al. Alignment over heterogeneous embeddings for question answering. NAACL'19.
- Yadav et. al. Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering. ACL'20.

Game-theoretic vocab selection via the Shapley value and Banzhaf index



Game-theoretic Vocabulary Selection via the Shapley Value and Banzhaf Index

Roma Patel, Marta Garnelo, Ian Gemp, Chris Dyer and Yoram Bachrach



Brown University, DeepMind

Motivation

- Goal:** Obtain a task-specific and semantically meaningful vocabulary for a task
- Approach:** An iterative algorithm that uses Shapley values to compute relevance scores for words
- Performance:** Evaluate in comparison to other heuristics (frequency and TF-IDF) on a range of different task structures

Algorithm

- Our algorithm compares power indices of words with respect to other words in the dataset
- The power index is approximated as the average marginal contribution of the word across the samples
- We use an approximation algorithm to compute this, since sampling all subsets is intractable

Algorithm 2 Shapley Vocabulary Selection

```

1: Inputs: NLP dataset  $D$  with full vocabulary  $V$ 
2: for each word  $w$  in  $V$  do
3:    $\phi_w \leftarrow 0$  (initialise Shapley value estimate)
4:   for  $i=1$  to  $S$  (number of sampled permutations do
5:      $\pi \leftarrow$  Random-Permutation( $V$ )
6:      $C_1 \leftarrow b(w, \pi)$  (predecessors of  $w$ )
7:      $C_2 \leftarrow C_1 \cup \{w\}$  (predecessors including  $w$ )
8:      $f_{\theta}^{C_1} \leftarrow$  TrainModel( $C_1$ ) (Train on vocabulary  $C_1$ )
9:      $f_{\theta}^{C_2} \leftarrow$  TrainModel( $C_2$ ) (Train on vocabulary  $C_2$ )
10:     $m(w, \pi) \leftarrow q(f_{\theta}^{C_2}) - q(f_{\theta}^{C_1})$ 
11:     $\phi_w \leftarrow \phi_w + m(w, \pi)$ 
12:   end for
13:    $\phi_w \leftarrow \frac{1}{S} \phi_w$  (average marginal contributions)
14: end for
15: Rank words in  $V$  based on Shapley estimates  $\pi_w$ 
16: Return top  $k$  words in ranking

```

Vocab size vs Performance

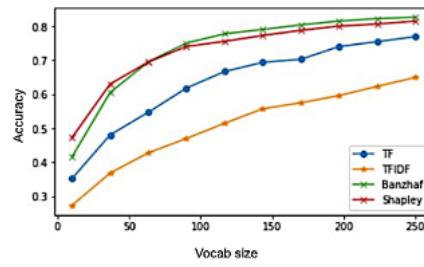
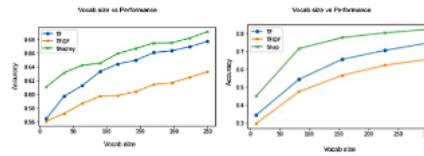


Figure 1: Comparing Shapley, Banzhaf, TF (term frequency) and TF-IDF on AG-news (a document classification task). We see that both game-theoretic algorithms (red and green) outperform the other heuristics for all vocabulary sizes.

Evaluation

Task & Dataset	Method	Vocab	Acc
SST-2 (Socher et al., 2013)	TF-IDF	17,539	80.2
	Frequency		80.3
	Banzhaf		81.7
	Shapley		81.9
COLA (Warstadt et al., 2019)	TF-IDF	9007	63.5
	Frequency		63.7
	Banzhaf		63.9
	Shapley		64.2
SNLI (Bowman et al., 2015b)	TF-IDF	42,392	83.9
	Frequency		83.9
	Banzhaf		84.1
	Shapley		84.3
QQP (Wang et al., 2018)	TF-IDF	117,303	80.8
	Frequency		81.2
	Banzhaf		81.9
	Shapley		81.9
AG-NEWS (Zhang et al., 2015)	TF-IDF	159,697	79.6
	Frequency		78.5
	Banzhaf		79.9
	Shapley		80.2
YELP (Zhang et al., 2015)	TF-IDF	458,705	84.5
	Frequency		83.9
	Banzhaf		86.7
	Shapley		87

Table 1: Performance of vocabulary selection methods across datasets and tasks, at a target vocabulary size of $|V'| = 750$ words (column 3 is initial vocabulary size). Note performance is lower than state-of-the-art methods, as results are based on a significantly reduced vocabulary size (and using a simple LSTM architecture, with no hyperparameter tuning).



A flexible natural language interface for web navigation



Motivation & Challenges

All assistants have started executing user tasks by directly interacting with the web. User commands are mapped into instructions that a browser can execute.

Problem:

- Existing approaches map commands directly into **low-level UI actions**, which is effective only in controlled or single-application environments.
- Websites are constantly updated, and users may want to execute the same task in any site of their choice, thus requiring constant model re-training.

Can we build a **flexible web navigation system** that does not require building website-specific models and can scale across websites?

Instead of low-level UI actions, we map commands into **concept-level actions** expressing what a user perceives when glancing at a website UI.

Learning concept-level actions can lead to a more **flexible NL interface** for web navigation.

Challenges

- Existing systems do not work well with environments that have a fixed set of known actions, which is not the case with real websites.
- The same concept-level action can have **different lexical representations and parameter schemas** across websites.
- In-domain websites can support **different actions** that change over time.

FLIN Key Insights

Semantic Parsing via Ranking

Leveraging the semantics of the symbols (name of action, parameter set and parameter values) in the logical form (navigation instruction) to learn how to match the given command with the most relevant navigation instruction.

C: Find me an Italian restaurant for me and my friend at 7 pm.

Time: 7:00 PM ≈ People: me and my friend ≈ Time: 7 pm ≈ Search: Italian

Action intent: find ≈ Let's go

Two sub-tasks:

1. Action recognition
2. Parameter recognition and value assignment

Find your table for any occasion

Closed-domain parameters of "let's go": Time, place, date

Open-domain parameter of "let's go": place

Parameter names may provide semantic categories which can act as semantic clues.

Evaluation Setup

Website (Domain)	# Pg	# Art	# Par	Train / Valid / Test
opentable.com (R)	8	56	56	14322 / 2864 / 1911
hotels.com (H)	7	54	19	1995 / 387 / 1587
bookair.co.uk (R)	7	54	19	16687 / 3240 / 2240
lyft.com (D)	1	77	46	7 / 1120 / 1120
zomato.com (D)	11	52	40	1687 / 1299 / 953
ebay.com (D)	11	52	40	1700 / 1299 / 953
amazon.com (S)	11	56	37	17 / 1120 / 1120

Experimental Results

Action accuracy	Acc # of param.	F1	EMA	PA-100	200% precision accuracy	FLIN and its variants adapt well to previously-unseen websites						
R: opentable.com (training website)	0.935	0.699	0.306	0.310	0.949	0.485	0.321	0.380	0.854	0.415	0.318	0.379
R: opentable.com (H)	0.935	0.699	0.306	0.310	0.949	0.485	0.321	0.380	0.854	0.415	0.318	0.379
H: hotels.com (training website)	0.937	0.681	0.479	0.506	0.926	0.481	0.320	0.356	0.824	0.412	0.322	0.363
H: hotels.com (D)	0.939	0.674	0.463	0.509	0.920	0.479	0.319	0.356	0.824	0.412	0.322	0.363
D: zomato.com (training website)	0.933	0.708	0.413	0.469	0.906	0.505	0.154	0.280	0.883	0.426	0.221	0.309
D: zomato.com (D)	0.939	0.674	0.463	0.509	0.920	0.479	0.319	0.356	0.824	0.412	0.322	0.363
S: ebay.com (D)	0.932	0.684	0.466	0.502	0.919	0.487	0.317	0.354	0.817	0.408	0.214	0.318
S: amazon.com (S)	0.932	0.684	0.466	0.502	0.919	0.487	0.317	0.354	0.817	0.408	0.214	0.318

FLIN shows robustness in the face of new commands from the DialQueries dataset achieving over 50% A-acc and 46% PA-100.

Performance of FLIN on DialQueries dataset

Time Type	8:00	9:00	9:30	10:00
Action no-predict	0.77	0.89	0.93	0.93
Action no-resolution	0.77	0.88	0.90	0.90
Predict identify slot	0.10	0.77	0.31	0.31
domain parameter	0.11	0.77	0.44	0.44
Character parameter	0.11	0.77	0.44	0.44
Full to extract open-	0.55	1.11	0.94	0.70
domain parameter val	0.55	1.11	0.94	0.70

Error analysis

Code and dataset are available at: <https://github.com/microsoft/flin-n2web>

Papers in discourse & pragmatics

Session 5B

Bridging anaphora resolution: making sense of the SOTA

Task

Bridging resolution aims to identify and resolve context-dependent but non-identical mentions

Even if *baseball* triggers losses at CBS - and he doesn't think it will - "I'd rather see *the games* on our air than on NBC and ABC," he says.

Existing Approaches

- Rule-based (Hou et al., 2014, Rosiger 2018)
 - Available training corpora may be too small to train a complex model
 - Designed 18 rules in total; rulesets for different corpora are slightly different
- Learning-based: Neural resolver by Yu and Poesio (2020) [Current state of the art]
 - Multi-task learning (MTL) of entity coreference and bridging resolution

Goal

- Understand the state of the art by answering two questions:
 - How is the MTL approach better than its rule-based counterparts?
 - What needs to be improved in MTL?

Evaluation Setup

- Corpora:** ISNotes (50 WSJ news articles, 663 anaphors), BASHI (50 WSJ news articles, 459 anaphors), ARRAU RST (413 news docs, 3777 anaphors)
- Setting:** Full bridging resolution
 - Input: gold mentions
 - Tasks: 1) Recognition: identify bridging anaphors, 2) Resolution: resolve them to their antecedents
- Evaluation metrics:** Precision, recall, and F-score for recognition and resolution

How is MTL better than rule-based approaches?

- We propose a **hybrid approach** to bridging resolution
 - A pipeline system, where we first apply the hand-crafted rules to identify bridging links, and then employ the MTL-based model to resolve any anaphoric mentions that are not resolved by the rules

Rules → **MTL**

If Hybrid outperforms Rules and MTL, then these two approaches have **different strengths and weaknesses** and should be viewed as **complementary rather than competing approaches**

Results: Recognition & Resolution Recall

- Hybrid's recalls are substantially higher than those of Rules and MTL for recognition and resolution
 - Rules and MTL make different mistakes (i.e., they complement each other's weaknesses)

	ISNotes	DAG-E	ARRAU RST
Recognition	~18	~22	~20
Resolution	~12	~20	~18

Results: Recognition & Resolution F-scores

- Hybrid achieves the state of the art results on all three datasets
- On ARRAU RST, performances of Hybrid and MTL are very close. Unlike in ISNotes and BASHI, where Rules's precision is higher than MTL's, in ARRAU RST, Rules's precision are more or less at the same level as MTL's

	ISNotes	DAG-E	ARRAU RST
Recognition	Rules: ~45, MTL: ~48, Hybrid: ~50	Rules: ~45, MTL: ~48, Hybrid: ~50	Rules: ~45, MTL: ~48, Hybrid: ~50
Resolution	Rules: ~25, MTL: ~28, Hybrid: ~28	Rules: ~25, MTL: ~28, Hybrid: ~28	Rules: ~25, MTL: ~28, Hybrid: ~28

Results: Rules and MTL on each rule category

	ISNotes	BASHI	ARRAU RST
Recognition	Rules: ~100, MTL: ~80	Rules: ~100, MTL: ~80	Rules: ~100, MTL: ~80
Resolution	Rules: ~20, MTL: ~15	Rules: ~20, MTL: ~15	Rules: ~20, MTL: ~15

Observation: Number of gold anaphors that satisfy a rule condition is **smaller** in BASHI, whereas number of gold mentions that satisfy an anaphor condition is larger in BASHI

- BASHI has longer documents, which could explain why more gold mentions satisfy anaphor conditions
- Some cases of bridging are not annotated in BASHI. (e.g., *Folk doctors* also prescribe it for kidney, bladder and urethra problems, duodenal ulcers and hemorrhoids. *Some* apply it to gouty joints.)

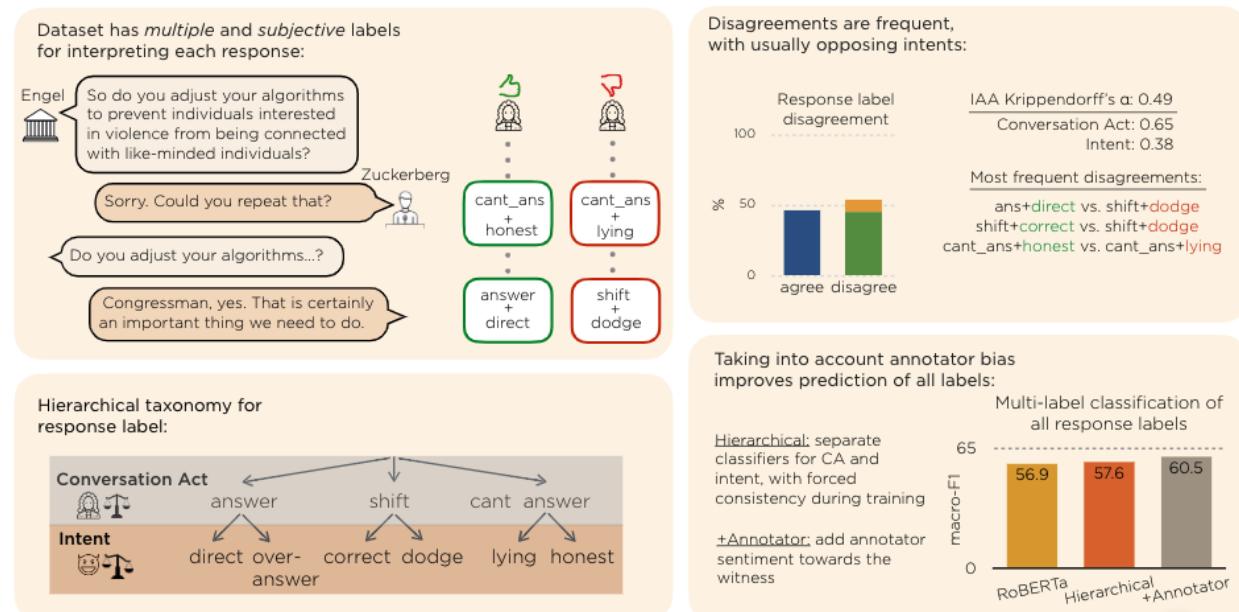
Error Analysis: What needs to be improved in MTL?

- Coreference anaphor [Recognition Precision Errors; 14-30%]**
 - Occurs when a gold coreference anaphor is misclassified as a bridging anaphor
 - Example: After three Sagos were stolen from his home in Garden Grove, "I put *a big iron stake* in *the ground* and tied the tree to *the stake* with a chain, " he says proudly.
 - MTL makes these mistakes because it is trained on coreference and bridging in the multi-task setting
- Indefinite expression [Recognition Recall Errors; 48-71%]**
 - Occurs when a system misclassifies an indefinite bridging anaphor as a NEW mention
 - Example: Currently, *Boeing* has a backlog of about \$80 billion, but *production* has been slowed by a strike of 55,000 machinists, which entered its 22nd day today.
 - Syntactic forms of many NEW instances and indefinite bridging anaphors are the same. Thus, it is not easy for model to distinguish between them
- Unmodified expression [Resolution Precision Errors; 23-63%]**
 - Occurs when a predicted anaphor is a short mention without modifiers (e.g., their imprisonment)
 - Such a mention is semantically less rich and is therefore harder to resolve

Did they answer? Subjective acts and intents in conversational discourse

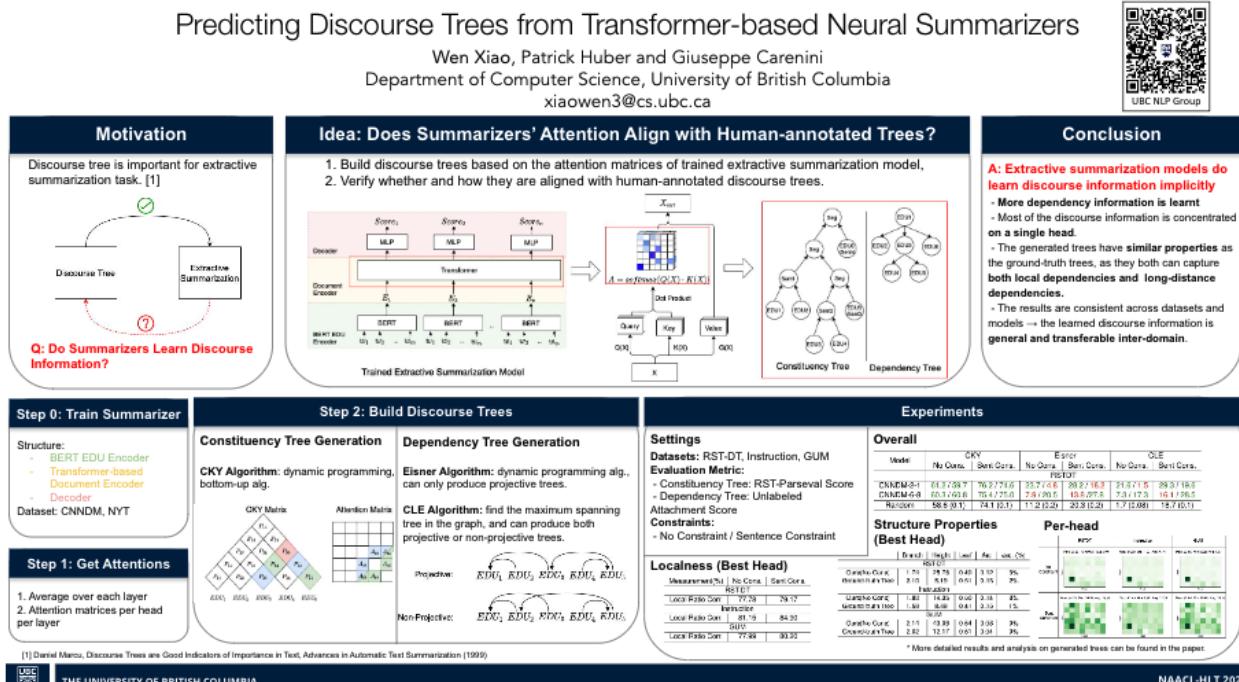
NAACL 2021 papers

51



Session 12A

Predicting discourse trees from Transformer-based neural summarizers



Is incoherence surprising? Targeted evaluation of coherence prediction from LMs



What Constitutes Coherence?

The cashier was counting the dollar bills at her desk.
Two men rushed into the store and held their guns up.
Everyone panicked and started to scream.
The men threatened the people to remain quiet.
The cashier handed them the cash so they would go away.

Figure 1: Text from ROCStories with colour-coded coherence relations

how are you ? being an old man , i am slowing down these days
hi , my dad is old as well , they live close to me and i see them often
that is a great thing honor your dad with your presence
sure , i pick him up for church every sunday with my ford pickup
sounds wonderful my wheelchair can go very fast on various terrains
i guess that means you do not go hunting often ? i love hunting , i own 3 guns

Figure 2: Dialogue extract from PERSONCHAT with colour-coded coherence relations

Dialogue: DIALoGPT (Zhang et al. 2020)

- GPT-2 fine-tuned on Reddit data, integrating speaker change
- Used by Mehri & Eskenazi (2020) to evaluate other dialogue models based on scores for hand-crafted follow-up utterances (e.g. “How interesting!” vs. “That is boring.”)
- Human evaluation revealed **coherence** to be among the most important factors of overall dialogue quality, but unclear whether and how this notion is represented in the model

Test Suites and Results

→ See paper for details on integration of existing test suites for Sentence Shuffling, Story Cloze and Winograd Schema

Referring Expressions

Minimal pairs constructed from ARRAU corpus (Uryupina et al., 2020):

context: and there's a ladder coming out of the tree and
there's a man at the top of the ladder
original: you can't see *him* yet
perturbed: you can't see *the man at the top of the ladder* yet

Results:

	wsj	vpc	Dialogue	Fiction
GPT-2	0.53	0.56	0.47	0.42
DIALoGPT	0.44	0.51	0.47	0.36

Previous Work

Coherence Evaluation: Sentence Shuffling (Barzilay & Lapata, 2008)



Figure 3: Which notions of coherence does random shuffling break exactly?

Targeted Syntactic Evaluation (Marvin & Linzen 2018)

Condition	Regions						continuation
	intro	vp[subj]	prep	the	prep[vp]	matrix verb	
match sing	The farmer	near	the	clerk	knows	many people	
mismatch sing	The farmer	near	the	clerk	knows	many people	
overmatch plural	The farmers	near	the	clerk	knows	many people	
match plural	The farmers	near	the	clerk	knows	many people	Item 1
matchsing	The manager	to the side of	the	architect	likes	to gamble	
mismatchsing	The manager	to the side of	the	architect	likes	to gamble	
overmatchplural	The manager	to the side of	the	architect	likes	to gamble	
matchplural	The managers	to the side of	the	architect	likes	to gamble	Item 2

Prediction: (matchsing matrix verb < mismatchsing matrix verb)
& (matchplural matrix verb < mismatchplural matrix verb)

Figure 4: SyntaxGym (Gauthier et al. 2020): Framework for syntactic evaluation of language models using minimal pairs and predictions evaluated on model's surprisal scores

CoherenceGym							
• Extend SyntaxGym framework to phenomena beyond syntax							
• Experiment with							
– Integrating existing test suites							
– Automatic creation of test suites from existing corpora							
• Evaluate pre-trained language and dialogue models							
• Coherence Detection Score: Proportion of items for which incoherent version is more surprising than coherent counterpart							
Pre-trained Models							

Discourse: GPT-2 (Radford et al. 2019)

- Shown to perform well on down-stream tasks that require some notion of coherence, such as story generation (See et al. 2019), but automatic measures only suited to evaluate diversity, better methods needed to actually measure **text coherence**

Probing for bridging inference in Transformer LMs

Explicit Connectives

Minimal pairs constructed from Disco-Annotation corpus (Popescu-Belis et al., 2012):

context: I am sure this Parliament will respond enthusiastically to this news
original: **as**
perturbed: **though**
continuation: It is exactly what we were pressing for.

Results:

GPT-2							
Connective used in manipulation							
CONNECTIVE SENSE	although	as	however	since	though	while	
as_causal	0.44	–	0.89	0.28	0.64	0.72	0.60
as_circumlocution	0.06	–	0.63	0.67	0.77	0.77	0.77
as_experiential	0.33	–	0.87	0.67	0.33	1.00	1.00
as_preposition	0.99	–	0.99	0.98	0.99	0.99	0.99
as_temporal	0.95	–	0.95	0.86	1.00	0.81	0.95

Speaker Commitment

Minimal pairs constructed from contradictions in DialogueNLI corpus (Welleck et al., 2019):

context: At since the beginning of the year, I am a nurse.
original: **I** am a kindergarten teacher.
perturbed: **A**: I am a kindergarten teacher.

Results:

contradiction	
DIALOGPT	0.59

Conclusion

- Coherence is more nuanced than sentence order shuffling can reflect
- Some notions of coherence seem to be encoded in these models, others are not detectable
- ⇒ **Targeted evaluation can help in guiding model improvements**

Next Steps

- Creating more high quality test suites
 - Templating approaches for test suite creation
 - Human evaluation as baselines
- Adding more models (impact of different sizes/architectures)
- Investigating extensions for different languages

Resources

Paper: <https://arxiv.org/abs/2105.03495>
Code: <https://github.com/AnneBeyer/coherencegym>

PROBING FOR BRIDGING INFERENCE IN TRANSFORMER LANGUAGE MODELS

Onkar Pandit¹ and Yufang Hou²

¹INRIA, Lille, France [✉ onkar.pandit@inria.fr]

²IBM Research, Dublin, Ireland [✉ yhou@ie.ibm.com]

MAIN CONTRIBUTIONS	
<ul style="list-style-type: none"> Investigated inner working of pre-trained transformer based language models, specifically for Bridging information. Presented two approaches of investigation: <ul style="list-style-type: none"> Probing of individual attention heads Of-cloze task to examine whole model 	

FINDINGS	
<ul style="list-style-type: none"> Pre-trained models capture substantial bridging information. Pre-trained ROBERTA-Large model achieved 28.05% accuracy for bridging which is comparable with state-of-the-art BARQA [1]. Higher layer capture more bridging information compared to middle or lower layers. This finding is in-line with previous findings, complex linguistic information is captured by higher layers. BERT fails at capturing sophisticated common sense information required to resolve some bridging pairs. Further, it also fails at resolving some pairs that require long contexts. 	

PROBING OF INDIVIDUAL ATTENTION HEADS

- Attention heads are crucial part of transformer models.
- We measure bridging signal captured by each attention head.
- We consider bridging signal from anaphor to antecedent as well as from antecedent to anaphor.
- Anaphor to antecedent bridging signal is calculated as ratio of attention given to antecedent by anaphor to cumulative attention paid to all the tokens. Similarly, antecedent to anaphor bridging signal is measured.

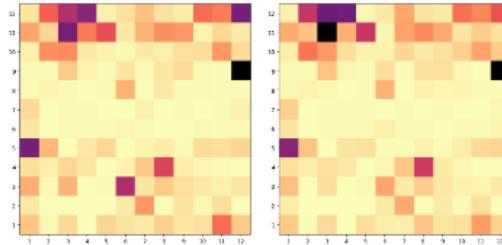


Figure 1:Bridging signals with BERT-base-cased model where anaphor and antecedents are 2 sentences apart. Bridging signals from anaphor to antecedent are shown in the first heatmap and the reverse signals in the second. In both heatmaps, the x-axis shows the attention head number and the y-axis shows the layer number.

Observations:

- Higher layers capture more bridging signal.
- Specific attention heads such 5:1, 9:12, 11:3, and 12:2:4, consistently capture bridging signal.
- Distance between anaphor-antecedent grows, bridging signal weakens

OF-CLOZE TASK INVESTIGATION

- This is to inspect whole model; complementary to previous approach.
- Exploit syntactic structure: Anaphor of Antecedent encodes many bridging relations.
- Fill-in-the-blank formulation : “[context] Anaphor of [MASK]” and predict [MASK] token with BERT
- Candidate antecedents for [MASK]: previous mentions and select the highest scoring candidate as prediction
- Of-cloze test context: [...] 22% of the firms said employees or owners had been robbed on their way to or from work or while on the job. **Seventeen percent of [MASK]** [...].

Experimental Set-up:

- Candidate scope –**
 - Salient/local mentions as candidate antecedents – Salient mentions – mentions from the first sentence of the document.
 - Local mentions – sentence containing anaphor and previous 2 sentences.
 - All previous mentions occurring before anaphor
- Context scope –**
 - Different contexts
 - Only Anaphor : “anaphor of [MASK]” without any context
 - Anaphor sentence: “(sentence)anaphor of [MASK]”
 - Anaphor + antecedent sentence: “[antecedent sentence] +(sentence)anaphor of [MASK]”
 - More context: “first sentence of the document + prev. two sentences + (sentence)anaphor of [MASK]”
 - Removing Of from context.
 - Perturbed context.

RESULTS

Antecedent Candidate Scope	BERT-Base	BERT-Large
<i>Prominent attention heads</i>		
(1) Salient/nearby mentions	20.15	-
<i>Of-Cloze Test</i>		
(2) Salient/nearby mentions	31.64	33.71
(3) All previous mentions	26.36	28.78
<i>Of-Cloze Test: Ante. in the provided contexts</i>		
(4) All previous mentions	29.00	30.88
<i>Of-Cloze Test: Ante. out of the contexts</i>		
(5) All previous mentions	10.98	16.48

Table 1:Result of selecting antecedents for anaphors with two different probing approaches (Prominent attention heads and Of-Cloze Test)

Context Scope with perturb	without
only anaphor	17.20 5.62
ana sent.	22.82 7.71
ana+ante sent.	27.81 9.61
more context	26.36 12.21
	11.41

Table 2:Accuracy of selecting antecedents with different types of context using BERT-of-Cloze Test

Distance	Accuracy
salient*	38.65
0	26.92
1	20.58
2	17.30
>2	10.98

Table 3:Anaphor-antecedent distance-wise accuracy with the BERT-base-cased model. * indicates that the antecedent is in the first sentence of the document.

REFERENCES

- [1] Yufang Hou.
Bridging anaphora resolution as question answering.
In *ACL 20*.

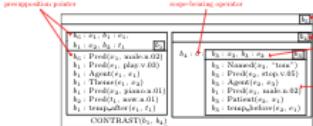
  paper

Universal Discourse Representation Structure Parsing
 Jiangming Liu, Shay B. Cohen, Mirella Lapata
 and Johan Bos
 University of Edinburgh and University of Groningen

 code 

Discourse Representation Theory

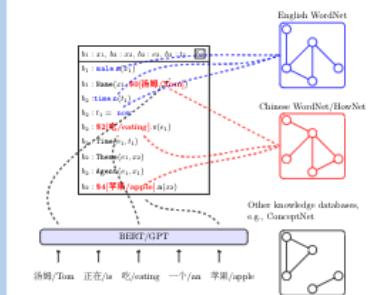
- Discourse Representation Theory (DRT; Kamp, 1981; Kamp and Reyle, 1993) developed to deal with multiple linguistic phenomena, such as predicate-argument, coreference, scope, quantification, presupposition, tense and aspect, and inter-sentence semantics, and analyze the meanings of texts.
- DRT uses Discourse Representation Structures (DRSs) to represent a hearer's mental representation of a discourse as it unfolds over time.
- It can be converted into the first-order logic form with simple rules.
- Compared to other semantic representations, DRT is able to represent more linguistic phenomena within and across sentences.



The main is going to play the piano. Tom might stop him.

Characteristics and Advantages in Universal DRS

- Link to knowledge bases and link to language models
- It bridges contextual information to knowledge bases under the semantic logics.



Cross-lingual Approach

- Machine translation system
- Word alignments between English and other languages are required.
- One-to-many approach translates gold-standard English (training data) to non-English text and trains multiple parsers

Universal Discourse Representation Structures

- Ground the symbols to the words in texts
- Remove the language-specific word senses



► Capture semantic patterns within and across languages

Low-resource language Experiments on the Parallel Meaning Bank

DRS parsing results



Universal DRS parsing results



► We construct Universal DRS dataset for **more than 100** languages

The constructed Universal DRS quality

Language	BLEU	F1
de	65.01	94.21
it	61.22	94.41
el	69.12	94.06
avg	65.12 (+3.9)	92.23 (+1.98)

Where the errors come from?

- Translation errors
- Translation divergences (Dorr, 1994): promotional, demotional, structural, conflational, lexical, categorical, and thematic
- Word alignment errors

	de	it	el	all
correct	11	10	7	7
translation-error	0	0	4	4
translation-divergence	1	0	0	2
alignment-error	4	4	3	3

informatics ilcc
 University of Edinburgh, University of Groningen <http://www.ilcc.inf.ed.ac.uk/> <https://www.rug.nl/>

Decontextualization: making sentences stand-alone

Decontextualization: Making Sentences Stand Alone

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm,
Tom Kwiatkowski, Dipanjan Das, Michael Collins



Data /
Demo at
GitHub!



Overview

We isolate and define the problem of sentence decontextualization: taking a sentence together with its context and rewriting it to be interpretable out of context, while preserving its meaning.

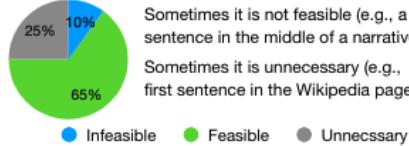
This can support models for question answering, dialogue agents, and summarization often interpret the meaning of a sentence in a rich context and use that meaning in a new context.

Decontextualization Examples

Document Title: Croatia at the FIFA World Cup	
Original	Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.
Decontextualized	The Croatia national football team's best result in FIFA world cup thus far was reaching the 2018 final, where they lost 4-2 to France.
Original	It stars professional dancer Lauren Taft alongside Petrica .
Decontextualized	The music video for Shut Up and Dance stars professional dancer Lauren Taft alongside Nicholas Petrica .

Task

Given a Wikipedia page and a sentence inside it, rewrite the sentence such that it can stand alone.



Sometimes it is not feasible (e.g., a sentence in the middle of a narrative).

Sometimes it is unnecessary (e.g., first sentence in the Wikipedia page)

Dataset

Sentences from English Wikipedia, specifically answer sentence from Natural Questions dataset and another sentence sampled from the same document.

Train dataset 1-way annotated, evaluation dataset 5-way annotated.

	Train	Dev	Test
# examples	11,290	1,945	1,945

Edits required to decontextualize

Phenomena	Example	% examples with phenomena
Pronoun / NP Swaps	he Bernie Sanders	40%
Bridging	characters characters of Toy Story 3	19%
Name / Acronym Expansion	Clinton Hillary Clinton	11.5%
Add Information	Charles Darwin Charles Darwin, an English naturalist and biologist	10%
Discourse Marker removal	However	3.5%

Original In 1850 , the first experimental electric telegraph line was started between Calcutta and Diamond Harbour.

Decontextualized In 1850, the first experimental electric telegraph line in India was started between Calcutta and Diamond Harbour.

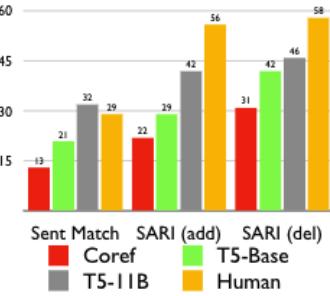
Original Although offered reinstatement after the threat is over, Hobbs decides to remain officially retired to spend more time with his daughter and his new "family", being Dom's team.

Impossible!

Intrinsic Evaluation

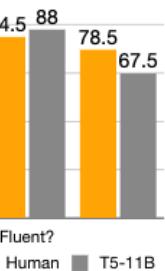
Evaluation Measures	Exact Match (on one of four references, feasible examples only)	
	SARI-add (added token overlap F1)	SARI-del (deleted token overlap F1)
Models	Coref SpanBERT model, Untrained baseline	
	Base line	
	BART-small Seq2Seq model, finetuned with decontextualized data	
Upper Bounds	BART-large	
	Human Annotator	

Results



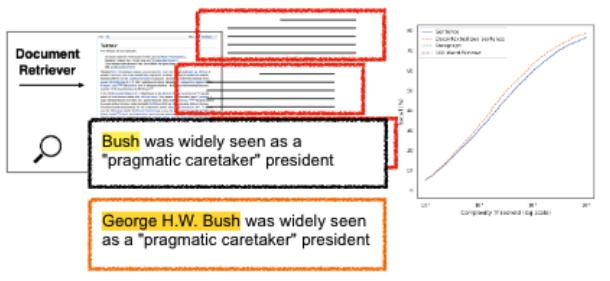
- Well trained seq2seq model performs surprisingly well
- Vanilla evaluation metric is challenging to tell apart good vs. bad decontextualization

Manual Eval



Application 1: Decontextualization as preprocessing

Use decontextualization to generate retrieval corpus for open domain QA.
Using decontextualized sentences as a retrieval corpus provides a better retrieval performance than using the original sentence.



Application 2: Decontextualization As Is

For question answering, instead of showing answer highlighted in the original answer sentence, we present an answer highlighted in a decontextualized answer sentence.

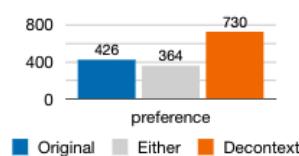
Q: what is the most viewed video on youtube in 24 hours?

Original

Decontextualized

The most viewed music video in this time period is Taylor Swift 's Look What You Made Me Do.

The most viewed music video on Youtube within 24 hours of its release is Taylor Swift 's Look What You Made Me Do.



Users preferred decontextualized answer sentence, for providing succinct yet informative answer to the question.

Papers in ML4NLP

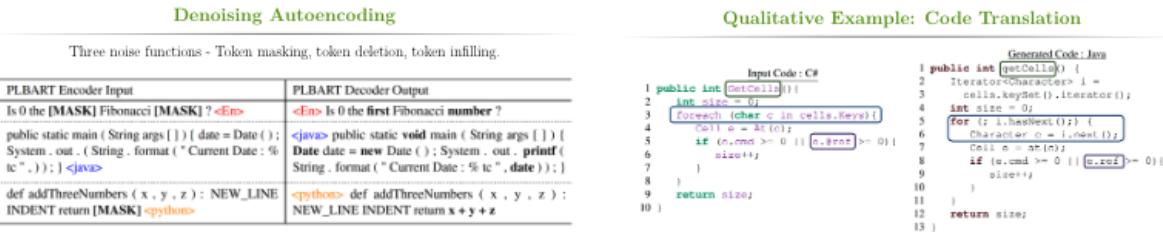
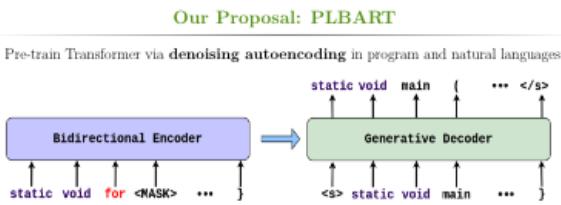
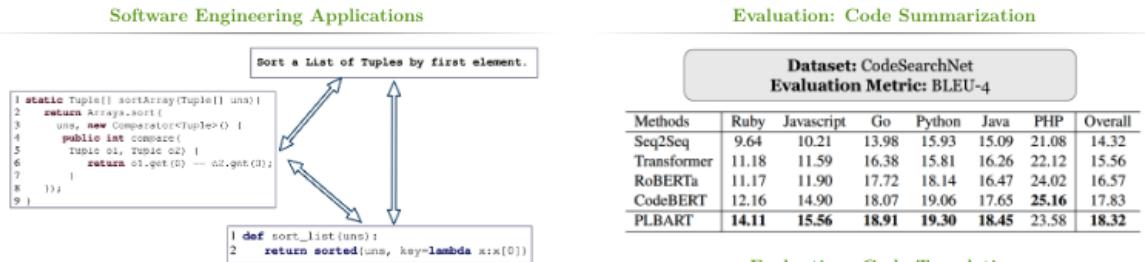
Session 8A

Unified pre-training for program understanding and generation

Unified Pre-training for Program Understanding and Generation

Wasi Uddin Ahmad^{†*}, Saikat Chakraborty^{‡*}, Baishakhi Ray[†], and Kai-Wei Chang[†]

[†]University of California, Los Angeles, [‡]Columbia University



How many data points is a prompt worth?



How many data points is a prompt worth?

Combining prompting and traditional supervision

Teven LE SCAO, Alexander M. RUSH
Hugging Face



Prompt-based methods have emerged as competition to standard fine-tuning. We combine those two paradigms and show that this yields an edge over traditional fine-tuning. We measure this advantage in terms of additional data points' worth of information provided by the prompt.

1/ Our method

Prompts are a way to turn a language model into a task-specific classifier. Most pretrained language models are trained with a token prediction objective. The usual pretrain-then-finetune method removes that token prediction head at finetuning time to **only use the internal representations of the model**.

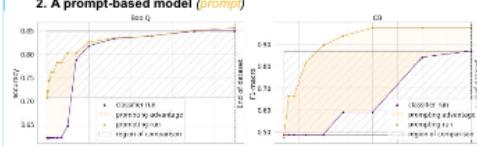
In contrast, a prompted model is used as a token predictor. It is presented with the input of the task and a **prompt: a short description of the task** that needs to be completed by a masked token. That token is then mapped into the desired class output by a **verbalizer**. (for example, Yes for 1 and No for 0).

Prompts are usually used in a zero-shot setting. We will **fine-tune** a prompted model using the **probability of the correct token as the loss objective**. This way, we can combine the information from the task description and supervised data.

2/ Results

We compare, on SuperGLUE and MNLI:

1. A linear classifier-based model (**classifier head**)
2. A prompt-based model (**prompt**)



On all tasks except WIC, the **prompt** model beats or is on par with the linear classifier model at all data scales. This effect is more pronounced at the low-data end: the additional information provided by the prompt matters more.

- Prompts add information at fine-tuning time
- This can combine with supervised data
- They still provide a useful inductive bias even without any zero-shot capabilities

3/ Data advantage

For a certain level of performance, the **prompt** and **head** models will require different amounts of data to reach that performance. This is the data advantage. We integrate this over the whole curve to get the **average data advantage**: this is how many data points the prompt brings on average. Up to 3500 on MNLI!

	Average Advantage (# Training Points)							
	MNLI	BoolQ	CB	COPA	MulRQC*	RTE	WIC	WSC
P vs H	3506 ± 536	752 ± 45	90 ± 2	288 ± 242	384 ± 378	282 ± 34	424 ± 74	281 ± 137
P vs N	150 ± 252	299 ± 81	78 ± 2	-	74 ± 56	405 ± 68	-354 ± 166	-
N vs H	3355 ± 612	453 ± 90	12 ± 1	-	309 ± 320	-122 ± 62	-70 ± 160	-

4/ Zero-shot vs adaptation

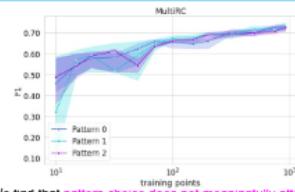
In order to study the zero-shot vs. adaptive nature of prompts, we introduce **null verbalizers**: models whose verbalizers are replaced with first names, so that zero-shot capability is at random chance. We compute which part of the data advantage is due to the zero-shot capability (**prompt** vs **null**) and which part to the inductive bias provided by the prompt (**head** vs **prompt**)

	MNLI	BoolQ	CB
P vs H	3506 ± 536	752 ± 46	90 ± 2
P vs N	150 ± 252	299 ± 81	78 ± 2
N vs H	3355 ± 612	453 ± 90	12 ± 1

	MulRQC*	RTE	WiC
P vs H	384 ± 378	282 ± 34	-424 ± 74
P vs N	74 ± 50	404 ± 68	-354 ± 166
N vs H	309 ± 320	-122 ± 62	-70 ± 160

We find that a significant part of this advantage is due to the inductive bias of the prompt, rather than zero-shot performance.

4/ Influence of pattern choice



We find that pattern choice does not meaningfully affect the results, as opposed to zero-shot learning.

Experimental setup

Testing on SuperGLUE + MNLI

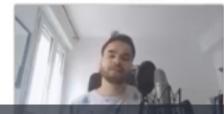
For every task, we fine-tune models on subsets of increasing data sizes
Best of 4 runs on every data size

Linear head model

- Start from RoBERTa-large
- Linear classification **head** instead of prediction head
- Fine-tuned via backpropagation on the predicted class
- Slight hyper-parameter tuning to be within 2 points of [SuperGLUE leaderboard](#)

Prompt model

- Start from RoBERTa-large
- Word prediction head with a prompt (3-4 different choices of prompts per task)
- Fine-tuned via backpropagation on the predicted token
- Reuses same hyperparameters as the **head** model.



A primer in BERTology: What we know about how BERT works?

A Primer in BERTology: What We Know about How BERT Works

Anna Rogers*, Olga Kovaleva†, Anna Rumshisky†

*University of Copenhagen | anna@cs.sdu.dk
†University of Massachusetts Lowell | olgakovleva, anna@ics.umass.edu

1. LOTS OF KNOWLEDGE CAN BE FOUND ✓

- It is possible to learn a linear transformation of BERT vector space that corresponds to syntactic trees (Hewitt et al. 2019)
- BERT's MLM prefers correct to incorrect verb forms (Goldberg 2019)
- The game that the guard hates [MASK] (Petroni et al. 2019)
- BERT has quite a lot of generic knowledge (Petroni et al. 2019)

2. KNOWLEDGE IS NOT USED ❌

- Input perturbations do not necessarily change predictions (Ettinger et al. 2020)
- The restaurant owner forgot which ~~purple~~ ~~the waiter~~ had [SERVED]
- BERT can't reason over the facts it "knows" (Forbes et al. 2019)
- If I can walk inside my house, I know that my house is bigger than I am
- BERT's knowledge sometimes comes from stereotypical associations (Petroni et al. 2019)

3. WHAT DOES PROBING SHOW? 😊

- Probing classifiers can extract lot of information from BERT embeddings about part of speech, syntactic chunks and roles, etc. (Li et al. 2019)
- Words sharing syntactic subtrees have larger impact on each other in MLM (Wu et al. 2020)

Issues with probing:

- different probing methods may lead to complementary or even contradicting conclusions (Wu et al. 2020)
- "the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used" (Tenney et al., 2019)

4. IS ATTENTION USEFUL? 😕

- Self-attention is intuitively appealing as a mechanism to encode syntactic relations (Clark et al. 2019)
- Possible to get the same results with attention patterns with other attention weights! (Jain et al. 2019)
- Most self-attention heads are not linguistically informative (Kovaleva et al. 2019)
- Heads surviving importance-based pruning are not necessarily linguistically informative (Prasanna et al. 2020)

5. LAYERS DIFFER 🎓

- Middle layers are the most transferable (Liu et al. 2019)
- Final layers are the most task-specific and the most affected by fine-tuning (Kovaleva et al. 2019)
- BERT may get "wiser" across layers (Tenney et al. 2019)

6. OVERPARAMETRIZED ❌

- Pre-training is not environmentally-friendly (Stavru et al. 2019)
- Most heads & layers can be pruned without much impact on performance (Michel et al. 2019, Kovaleva et al. 2019, Voita et al. 2019)
- 30-40% of the weights can be pruned without impact on downstream tasks (Gordon et al. 2020)
- Larger is not always better (Goldberg 2019, Lin et al. 2019)

7. ROOM FOR IMPROVEMENT 💡

- Some architectural choices can make BERT lighter
- BERT can be efficiently compressed
- Many proposals for improving the training process:
 - improving the training regimen (McCoy et al. 2019, ...)
 - tweaking the pre-training (Lin et al. 2020, ...)
 - tweaking the fine-tuning (Arsee et al. 2019, ...)
 - making it more stable (Mosbach et al. 2021)
 - ...

8. HIGH VARIANCE OF RUNS ✋

- Results vary a lot with fine-tuning initializations (Dodge et al. 2020)
- Some runs generalize better (McCoy et al. 2020)

- High variance still holds even if the majority of the initialized BERT weights are frozen (McCoy et al. 2019)
- Same data orders and initializations are better than others (Dodge et al. 2019)

9. IS IT BERT OR IS IT DATA? 😊

- Current benchmarks are too easy:
 - BERT learns shallow heuristics in NLU (McCoy et al. 2019, Zellers et al. 2019, Jin et al. 2020)
 - BERT learns shallow heuristics in GLUE (McCoy et al. 2019, Rogers et al. 2020, Sugawara et al. 2020)
 - Probably also elsewhere else
- BERT works pretty well even without pre-training (Kovaleva et al. 2019)
- Shallow heuristics can be used to reconstruct the model (Krishna et al. 2020)
- To be solved:
 - Better data (not to teach spurious correlations)
 - Better training (not to learn spurious correlations)
 - Better tests (to figure out whether it learned spurious correlations anyway)
 - What information is actually used at inference time? (amnesia probing (Elazar et al. 2020), pruning to interpret (Voita et al. 2019, Prasanna et al. 2020), ...)

This is a very brief overview. Check out our paper - it surveys over 150 studies! — +

Session 9E

Grouping words with semantic diversity

GROUPING WORDS WITH SEMANTIC DIVERSITY

Karine Chubarian² = Abdul Rafae Khan¹ = Anastasios Sidiropoulos² = Jia Xu¹ *

¹Department of Computer Science, Stevens Institute of Technology

²Department of Computer Science and Technology, University of Illinois at Chicago

* Authors ordered alphabetically

Introduction

- Problem: natural language inputs to NLP models are high-dimensional.
- open-vocabulary inputs inevitably bring rare and OOVs.
 - network complexity increases with input dimension.
- Q1: "Can we compute a generalized language representation to improve NLP applications?"
- word clustering as a many-to-one mapping.
 - to reduce the input vocabulary size and lower the input feature dimensions.
 - input information loss.



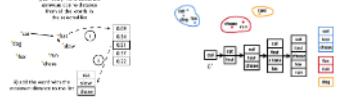
- Q2: "How can we design an algorithm that simplifies language representation while preserving meaning expressiveness?"
- the context of semantically diverse words varies more than that of semantically close words.
 - our diverse grouping uses context to distinguish words from the same group, leading to a more expressive representation.

Random Grouping

- randomly sample a phonetic group size that follows a Poisson distribution
 - uniform randomly sample K words and group
 - repeat for all groups
- random grouping does not guarantee semantic diversity in a group.

Distance-based Diverse Grouping

- randomly pick the i^{th} word and add to list L'
- compute each word's minimum cosine distances (MCD) to all words in list L'
- append the word with the maximum MCD to the list L'
- repeat steps 2 & 3 until all words ranked
- segment the ranked list into groups
- increase the distances among words in a group ignoring word frequencies.



Entropy-based Diverse Grouping

- consider the entropy with respect to a distribution induced by the relative frequencies of group unigrams.
- the entropy is maximal when the underlying distribution is close to uniform.
- minimize information loss
- adapt submodular maximization [1]
- grouping γ : a set of all pairs (w, γ_i) where $\gamma(w) = \gamma_i$.
- for each pair (w, γ_i) , perform one of the following if there is enough entropy gain.
 - put a word w into a group γ_i
 - remove a word w from a group γ_i
 - remove a word w from a group γ_i and then put another word v into a group γ_j (we allow either $w = v$ or $\gamma_i = \gamma_j$)
- assign ungrouped words to the group with smallest partial entropy.

Theorem

- given any precision parameter $\epsilon > 0$, our algorithm runs in polynomial time and is a $(1 - \frac{\epsilon}{\ln(1/\epsilon)})$ -approximation to the maximum unigram entropy.
- our algorithm is about $1/4$ away from optimal of maximizing the entropy.
- in typical case, our algorithm is very close to the optimal.
- our proof adapts [1] by showing grouping set family forms **matroid** and our objective function is submodular.

Matroid Properties

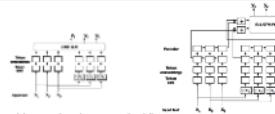
- Two Matroid properties:
- let Q define a grouping, then every $R \subset Q$ also defines a grouping.
 - let S, T define two groupings such that $|T| < |S|$, then there always exists a pair $(w, \gamma_i) \in S \setminus T$ such that when added to T results in a grouping.

Submodular Function

- to show that our mapping function is non-negative and submodular:
- let γ'_i, γ_j define two groupings introduced by Q, R respectively
 - every word w that can be added to γ'_i can also be added to γ_j .
 - $F_{w,i} / c_{w,i}$: word / group relative frequency
 - the entropy gain only depend upon partial entropies of group index i

$$-(c_{w,i} + F_w) \log(c_{w,i} + F_w) + c_{w,i} \log(c_{w,i}) - (c_{w,j} + F_w) \log(c_{w,j} + F_w) + c_{w,j} \log(c_{w,j})$$
 - thus $c_{w,i} < c_{w,j}$ where $c_{w,i}$ and $c_{w,j}$ is the relative frequency of group γ'_i and γ'_j
 - the entropy change is only for γ'_i and it is non-negative and monotone decreasing.
 - this implies, larger grouping gains less entropy than smaller grouping.

Combination Methods



To combine word and group embeddings:

- concatenation for Machine Translation and Language Modeling.
- linear combination for Part-of-Speech tagging.

Experimental Results

	IWSLT17 EN-FR	IWSLT17 ED
Baseline (Conv2DB)	17.6	22.85
Random Grouping	23.5	23.0
Poisson-based Random Grouping	23.0	23.6
Distance-based Grouping	23.6	21.99 (+5.76%)
Entropy-based Grouping	24.1 (+38.9%)	

	IWSLT17	MNTT18
Machine Translation (in BLEU%)	EN-DE DE-EN EN-ZH EN-FR	
Baseline (Conv2DB)	19.4 22.6 18.0 19.4	
Entropy-based Grouping	21.0 24.0 19.2 23.9 (+18.8%)	

Brown Corpus EN	
	Dev Test
Part-of-Speech Tagging	
Baseline (S LSTM)	5.32 5.39
Random Grouping	5.32 5.07
Poisson-based Random Grouping	5.56 5.27
Entropy-based Grouping	5.48 5.22 98.85 1.34 (+3.60%)

Acknowledgements

- National Science Foundation (NSF) Award No. 1747728
- National Science Foundation of China (NSFC) Award No. 61622524
- Google Cloud Research Program

References

- [1] Jon Lee et al. "Non-monotonic submodular maximization under matroid and knapsack constraints". In: Proceedings of the 41st Annual Symposium on Theory of Computing. (2009).

Modeling content and context with deep relational learning

Modeling Content and Context with Deep Relational Learning

Maria Leonor Pacheco and Dan Goldwasser
Department of Computer Science, Purdue University



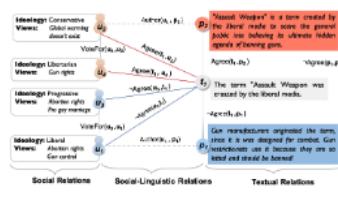
Motivation

- Goal:** Build natural language processing models that can
- Deal with noisy textual inputs
 - Model dependencies between different textual elements
 - Reason about the context that surrounds the language event

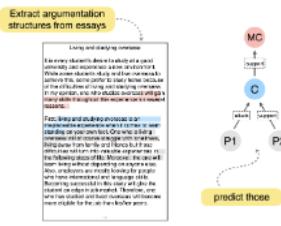
Contribution

- Introduce DRaIL, a **declarative neural-symbolic framework** designed to support a variety of NLP scenarios

Task 1: Debate Stance Prediction



Task 2: Argument Mining



Framework Overview

- Task defined by specifying **entities** and **relations**
 - Decision are defined using horn clauses
- $$\text{Reln}_1 \wedge \text{Reln}_2 \dots \text{Reln}_{n-1} \Rightarrow \text{Reln}_n$$
- DRaIL Program
- ```

rule_def:
 rule1_def:
 rule1_bdy: A1 > Agreec(A1)
 rule1_hbdy: B1 & C1
 rule1_rbdy: D1 & E1
 rule2_def:
 rule2_bdy: A2 > Agreec(A2)
 rule2_hbdy: B2 & C2
 rule2_rbdy: D2 & E2
 rule3_def:
 rule3_bdy: A3 > Agreec(A3)
 rule3_hbdy: B3 & C3
 rule3_rbdy: D3 & E3
 rule4_def:
 rule4_bdy: A4 > Agreec(A4)
 rule4_hbdy: B4 & C4
 rule4_rbdy: D4 & E4
 rule5_def:
 rule5_bdy: A5 > Agreec(A5)
 rule5_hbdy: B5 & C5
 rule5_rbdy: D5 & E5
 rule6_def:
 rule6_bdy: A6 > Agreec(A6)
 rule6_hbdy: B6 & C6
 rule6_rbdy: D6 & E6
 rule7_def:
 rule7_bdy: A7 > Agreec(A7)
 rule7_hbdy: B7 & C7
 rule7_rbdy: D7 & E7
 rule8_def:
 rule8_bdy: A8 > Agreec(A8)
 rule8_hbdy: B8 & C8
 rule8_rbdy: D8 & E8
 rule9_def:
 rule9_bdy: A9 > Agreec(A9)
 rule9_hbdy: B9 & C9
 rule9_rbdy: D9 & E9
 rule10_def:
 rule10_bdy: A10 > Agreec(A10)
 rule10_hbdy: B10 & C10
 rule10_rbdy: D10 & E10
 rule11_def:
 rule11_bdy: A11 > Agreec(A11)
 rule11_hbdy: B11 & C11
 rule11_rbdy: D11 & E11
 rule12_def:
 rule12_bdy: A12 > Agreec(A12)
 rule12_hbdy: B12 & C12
 rule12_rbdy: D12 & E12
 rule13_def:
 rule13_bdy: A13 > Agreec(A13)
 rule13_hbdy: B13 & C13
 rule13_rbdy: D13 & E13
 rule14_def:
 rule14_bdy: A14 > Agreec(A14)
 rule14_hbdy: B14 & C14
 rule14_rbdy: D14 & E14
 rule15_def:
 rule15_bdy: A15 > Agreec(A15)
 rule15_hbdy: B15 & C15
 rule15_rbdy: D15 & E15
 rule16_def:
 rule16_bdy: A16 > Agreec(A16)
 rule16_hbdy: B16 & C16
 rule16_rbdy: D16 & E16
 rule17_def:
 rule17_bdy: A17 > Agreec(A17)
 rule17_hbdy: B17 & C17
 rule17_rbdy: D17 & E17
 rule18_def:
 rule18_bdy: A18 > Agreec(A18)
 rule18_hbdy: B18 & C18
 rule18_rbdy: D18 & E18
 rule19_def:
 rule19_bdy: A19 > Agreec(A19)
 rule19_hbdy: B19 & C19
 rule19_rbdy: D19 & E19
 rule20_def:
 rule20_bdy: A20 > Agreec(A20)
 rule20_hbdy: B20 & C20
 rule20_rbdy: D20 & E20
 rule21_def:
 rule21_bdy: A21 > Agreec(A21)
 rule21_hbdy: B21 & C21
 rule21_rbdy: D21 & E21
 rule22_def:
 rule22_bdy: A22 > Agreec(A22)
 rule22_hbdy: B22 & C22
 rule22_rbdy: D22 & E22
 rule23_def:
 rule23_bdy: A23 > Agreec(A23)
 rule23_hbdy: B23 & C23
 rule23_rbdy: D23 & E23
 rule24_def:
 rule24_bdy: A24 > Agreec(A24)
 rule24_hbdy: B24 & C24
 rule24_rbdy: D24 & E24
 rule25_def:
 rule25_bdy: A25 > Agreec(A25)
 rule25_hbdy: B25 & C25
 rule25_rbdy: D25 & E25
 rule26_def:
 rule26_bdy: A26 > Agreec(A26)
 rule26_hbdy: B26 & C26
 rule26_rbdy: D26 & E26
 rule27_def:
 rule27_bdy: A27 > Agreec(A27)
 rule27_hbdy: B27 & C27
 rule27_rbdy: D27 & E27
 rule28_def:
 rule28_bdy: A28 > Agreec(A28)
 rule28_hbdy: B28 & C28
 rule28_rbdy: D28 & E28
 rule29_def:
 rule29_bdy: A29 > Agreec(A29)
 rule29_hbdy: B29 & C29
 rule29_rbdy: D29 & E29
 rule30_def:
 rule30_bdy: A30 > Agreec(A30)
 rule30_hbdy: B30 & C30
 rule30_rbdy: D30 & E30
 rule31_def:
 rule31_bdy: A31 > Agreec(A31)
 rule31_hbdy: B31 & C31
 rule31_rbdy: D31 & E31
 rule32_def:
 rule32_bdy: A32 > Agreec(A32)
 rule32_hbdy: B32 & C32
 rule32_rbdy: D32 & E32
 rule33_def:
 rule33_bdy: A33 > Agreec(A33)
 rule33_hbdy: B33 & C33
 rule33_rbdy: D33 & E33
 rule34_def:
 rule34_bdy: A34 > Agreec(A34)
 rule34_hbdy: B34 & C34
 rule34_rbdy: D34 & E34
 rule35_def:
 rule35_bdy: A35 > Agreec(A35)
 rule35_hbdy: B35 & C35
 rule35_rbdy: D35 & E35
 rule36_def:
 rule36_bdy: A36 > Agreec(A36)
 rule36_hbdy: B36 & C36
 rule36_rbdy: D36 & E36
 rule37_def:
 rule37_bdy: A37 > Agreec(A37)
 rule37_hbdy: B37 & C37
 rule37_rbdy: D37 & E37
 rule38_def:
 rule38_bdy: A38 > Agreec(A38)
 rule38_hbdy: B38 & C38
 rule38_rbdy: D38 & E38
 rule39_def:
 rule39_bdy: A39 > Agreec(A39)
 rule39_hbdy: B39 & C39
 rule39_rbdy: D39 & E39
 rule40_def:
 rule40_bdy: A40 > Agreec(A40)
 rule40_hbdy: B40 & C40
 rule40_rbdy: D40 & E40
 rule41_def:
 rule41_bdy: A41 > Agreec(A41)
 rule41_hbdy: B41 & C41
 rule41_rbdy: D41 & E41
 rule42_def:
 rule42_bdy: A42 > Agreec(A42)
 rule42_hbdy: B42 & C42
 rule42_rbdy: D42 & E42
 rule43_def:
 rule43_bdy: A43 > Agreec(A43)
 rule43_hbdy: B43 & C43
 rule43_rbdy: D43 & E43
 rule44_def:
 rule44_bdy: A44 > Agreec(A44)
 rule44_hbdy: B44 & C44
 rule44_rbdy: D44 & E44
 rule45_def:
 rule45_bdy: A45 > Agreec(A45)
 rule45_hbdy: B45 & C45
 rule45_rbdy: D45 & E45
 rule46_def:
 rule46_bdy: A46 > Agreec(A46)
 rule46_hbdy: B46 & C46
 rule46_rbdy: D46 & E46
 rule47_def:
 rule47_bdy: A47 > Agreec(A47)
 rule47_hbdy: B47 & C47
 rule47_rbdy: D47 & E47
 rule48_def:
 rule48_bdy: A48 > Agreec(A48)
 rule48_hbdy: B48 & C48
 rule48_rbdy: D48 & E48
 rule49_def:
 rule49_bdy: A49 > Agreec(A49)
 rule49_hbdy: B49 & C49
 rule49_rbdy: D49 & E49
 rule50_def:
 rule50_bdy: A50 > Agreec(A50)
 rule50_hbdy: B50 & C50
 rule50_rbdy: D50 & E50
 rule51_def:
 rule51_bdy: A51 > Agreec(A51)
 rule51_hbdy: B51 & C51
 rule51_rbdy: D51 & E51
 rule52_def:
 rule52_bdy: A52 > Agreec(A52)
 rule52_hbdy: B52 & C52
 rule52_rbdy: D52 & E52
 rule53_def:
 rule53_bdy: A53 > Agreec(A53)
 rule53_hbdy: B53 & C53
 rule53_rbdy: D53 & E53
 rule54_def:
 rule54_bdy: A54 > Agreec(A54)
 rule54_hbdy: B54 & C54
 rule54_rbdy: D54 & E54
 rule55_def:
 rule55_bdy: A55 > Agreec(A55)
 rule55_hbdy: B55 & C55
 rule55_rbdy: D55 & E55
 rule56_def:
 rule56_bdy: A56 > Agreec(A56)
 rule56_hbdy: B56 & C56
 rule56_rbdy: D56 & E56
 rule57_def:
 rule57_bdy: A57 > Agreec(A57)
 rule57_hbdy: B57 & C57
 rule57_rbdy: D57 & E57
 rule58_def:
 rule58_bdy: A58 > Agreec(A58)
 rule58_hbdy: B58 & C58
 rule58_rbdy: D58 & E58
 rule59_def:
 rule59_bdy: A59 > Agreec(A59)
 rule59_hbdy: B59 & C59
 rule59_rbdy: D59 & E59
 rule60_def:
 rule60_bdy: A60 > Agreec(A60)
 rule60_hbdy: B60 & C60
 rule60_rbdy: D60 & E60
 rule61_def:
 rule61_bdy: A61 > Agreec(A61)
 rule61_hbdy: B61 & C61
 rule61_rbdy: D61 & E61
 rule62_def:
 rule62_bdy: A62 > Agreec(A62)
 rule62_hbdy: B62 & C62
 rule62_rbdy: D62 & E62
 rule63_def:
 rule63_bdy: A63 > Agreec(A63)
 rule63_hbdy: B63 & C63
 rule63_rbdy: D63 & E63
 rule64_def:
 rule64_bdy: A64 > Agreec(A64)
 rule64_hbdy: B64 & C64
 rule64_rbdy: D64 & E64
 rule65_def:
 rule65_bdy: A65 > Agreec(A65)
 rule65_hbdy: B65 & C65
 rule65_rbdy: D65 & E65
 rule66_def:
 rule66_bdy: A66 > Agreec(A66)
 rule66_hbdy: B66 & C66
 rule66_rbdy: D66 & E66
 rule67_def:
 rule67_bdy: A67 > Agreec(A67)
 rule67_hbdy: B67 & C67
 rule67_rbdy: D67 & E67
 rule68_def:
 rule68_bdy: A68 > Agreec(A68)
 rule68_hbdy: B68 & C68
 rule68_rbdy: D68 & E68
 rule69_def:
 rule69_bdy: A69 > Agreec(A69)
 rule69_hbdy: B69 & C69
 rule69_rbdy: D69 & E69
 rule70_def:
 rule70_bdy: A70 > Agreec(A70)
 rule70_hbdy: B70 & C70
 rule70_rbdy: D70 & E70
 rule71_def:
 rule71_bdy: A71 > Agreec(A71)
 rule71_hbdy: B71 & C71
 rule71_rbdy: D71 & E71
 rule72_def:
 rule72_bdy: A72 > Agreec(A72)
 rule72_hbdy: B72 & C72
 rule72_rbdy: D72 & E72
 rule73_def:
 rule73_bdy: A73 > Agreec(A73)
 rule73_hbdy: B73 & C73
 rule73_rbdy: D73 & E73
 rule74_def:
 rule74_bdy: A74 > Agreec(A74)
 rule74_hbdy: B74 & C74
 rule74_rbdy: D74 & E74
 rule75_def:
 rule75_bdy: A75 > Agreec(A75)
 rule75_hbdy: B75 & C75
 rule75_rbdy: D75 & E75
 rule76_def:
 rule76_bdy: A76 > Agreec(A76)
 rule76_hbdy: B76 & C76
 rule76_rbdy: D76 & E76
 rule77_def:
 rule77_bdy: A77 > Agreec(A77)
 rule77_hbdy: B77 & C77
 rule77_rbdy: D77 & E77
 rule78_def:
 rule78_bdy: A78 > Agreec(A78)
 rule78_hbdy: B78 & C78
 rule78_rbdy: D78 & E78
 rule79_def:
 rule79_bdy: A79 > Agreec(A79)
 rule79_hbdy: B79 & C79
 rule79_rbdy: D79 & E79
 rule80_def:
 rule80_bdy: A80 > Agreec(A80)
 rule80_hbdy: B80 & C80
 rule80_rbdy: D80 & E80
 rule81_def:
 rule81_bdy: A81 > Agreec(A81)
 rule81_hbdy: B81 & C81
 rule81_rbdy: D81 & E81
 rule82_def:
 rule82_bdy: A82 > Agreec(A82)
 rule82_hbdy: B82 & C82
 rule82_rbdy: D82 & E82
 rule83_def:
 rule83_bdy: A83 > Agreec(A83)
 rule83_hbdy: B83 & C83
 rule83_rbdy: D83 & E83
 rule84_def:
 rule84_bdy: A84 > Agreec(A84)
 rule84_hbdy: B84 & C84
 rule84_rbdy: D84 & E84
 rule85_def:
 rule85_bdy: A85 > Agreec(A85)
 rule85_hbdy: B85 & C85
 rule85_rbdy: D85 & E85
 rule86_def:
 rule86_bdy: A86 > Agreec(A86)
 rule86_hbdy: B86 & C86
 rule86_rbdy: D86 & E86
 rule87_def:
 rule87_bdy: A87 > Agreec(A87)
 rule87_hbdy: B87 & C87
 rule87_rbdy: D87 & E87
 rule88_def:
 rule88_bdy: A88 > Agreec(A88)
 rule88_hbdy: B88 & C88
 rule88_rbdy: D88 & E88
 rule89_def:
 rule89_bdy: A89 > Agreec(A89)
 rule89_hbdy: B89 & C89
 rule89_rbdy: D89 & E89
 rule90_def:
 rule90_bdy: A90 > Agreec(A90)
 rule90_hbdy: B90 & C90
 rule90_rbdy: D90 & E90
 rule91_def:
 rule91_bdy: A91 > Agreec(A91)
 rule91_hbdy: B91 & C91
 rule91_rbdy: D91 & E91
 rule92_def:
 rule92_bdy: A92 > Agreec(A92)
 rule92_hbdy: B92 & C92
 rule92_rbdy: D92 & E92
 rule93_def:
 rule93_bdy: A93 > Agreec(A93)
 rule93_hbdy: B93 & C93
 rule93_rbdy: D93 & E93
 rule94_def:
 rule94_bdy: A94 > Agreec(A94)
 rule94_hbdy: B94 & C94
 rule94_rbdy: D94 & E94
 rule95_def:
 rule95_bdy: A95 > Agreec(A95)
 rule95_hbdy: B95 & C95
 rule95_rbdy: D95 & E95
 rule96_def:
 rule96_bdy: A96 > Agreec(A96)
 rule96_hbdy: B96 & C96
 rule96_rbdy: D96 & E96
 rule97_def:
 rule97_bdy: A97 > Agreec(A97)
 rule97_hbdy: B97 & C97
 rule97_rbdy: D97 & E97
 rule98_def:
 rule98_bdy: A98 > Agreec(A98)
 rule98_hbdy: B98 & C98
 rule98_rbdy: D98 & E98
 rule99_def:
 rule99_bdy: A99 > Agreec(A99)
 rule99_hbdy: B99 & C99
 rule99_rbdy: D99 & E99
 rule100_def:
 rule100_bdy: A100 > Agreec(A100)
 rule100_hbdy: B100 & C100
 rule100_rbdy: D100 & E100
 rule101_def:
 rule101_bdy: A101 > Agreec(A101)
 rule101_hbdy: B101 & C101
 rule101_rbdy: D101 & E101
 rule102_def:
 rule102_bdy: A102 > Agreec(A102)
 rule102_hbdy: B102 & C102
 rule102_rbdy: D102 & E102
 rule103_def:
 rule103_bdy: A103 > Agreec(A103)
 rule103_hbdy: B103 & C103
 rule103_rbdy: D103 & E103
 rule104_def:
 rule104_bdy: A104 > Agreec(A104)
 rule104_hbdy: B104 & C104
 rule104_rbdy: D104 & E104
 rule105_def:
 rule105_bdy: A105 > Agreec(A105)
 rule105_hbdy: B105 & C105
 rule105_rbdy: D105 & E105
 rule106_def:
 rule106_bdy: A106 > Agreec(A106)
 rule106_hbdy: B106 & C106
 rule106_rbdy: D106 & E106
 rule107_def:
 rule107_bdy: A107 > Agreec(A107)
 rule107_hbdy: B107 & C107
 rule107_rbdy: D107 & E107
 rule108_def:
 rule108_bdy: A108 > Agreec(A108)
 rule108_hbdy: B108 & C108
 rule108_rbdy: D108 & E108
 rule109_def:
 rule109_bdy: A109 > Agreec(A109)
 rule109_hbdy: B109 & C109
 rule109_rbdy: D109 & E109
 rule110_def:
 rule110_bdy: A110 > Agreec(A110)
 rule110_hbdy: B110 & C110
 rule110_rbdy: D110 & E110
 rule111_def:
 rule111_bdy: A111 > Agreec(A111)
 rule111_hbdy: B111 & C111
 rule111_rbdy: D111 & E111
 rule112_def:
 rule112_bdy: A112 > Agreec(A112)
 rule112_hbdy: B112 & C112
 rule112_rbdy: D112 & E112
 rule113_def:
 rule113_bdy: A113 > Agreec(A113)
 rule113_hbdy: B113 & C113
 rule113_rbdy: D113 & E113
 rule114_def:
 rule114_bdy: A114 > Agreec(A114)
 rule114_hbdy: B114 & C114
 rule114_rbdy: D114 & E114
 rule115_def:
 rule115_bdy: A115 > Agreec(A115)
 rule115_hbdy: B115 & C115
 rule115_rbdy: D115 & E115
 rule116_def:
 rule116_bdy: A116 > Agreec(A116)
 rule116_hbdy: B116 & C116
 rule116_rbdy: D116 & E116
 rule117_def:
 rule117_bdy: A117 > Agreec(A117)
 rule117_hbdy: B117 & C117
 rule117_rbdy: D117 & E117
 rule118_def:
 rule118_bdy: A118 > Agreec(A118)
 rule118_hbdy: B118 & C118
 rule118_rbdy: D118 & E118
 rule119_def:
 rule119_bdy: A119 > Agreec(A119)
 rule119_hbdy: B119 & C119
 rule119_rbdy: D119 & E119
 rule120_def:
 rule120_bdy: A120 > Agreec(A120)
 rule120_hbdy: B120 & C120
 rule120_rbdy: D120 & E120
 rule121_def:
 rule121_bdy: A121 > Agreec(A121)
 rule121_hbdy: B121 & C121
 rule121_rbdy: D121 & E121
 rule122_def:
 rule122_bdy: A122 > Agreec(A122)
 rule122_hbdy: B122 & C122
 rule122_rbdy: D122 & E122
 rule123_def:
 rule123_bdy: A123 > Agreec(A123)
 rule123_hbdy: B123 & C123
 rule123_rbdy: D123 & E123
 rule124_def:
 rule124_bdy: A124 > Agreec(A124)
 rule124_hbdy: B124 & C124
 rule124_rbdy: D124 & E124
 rule125_def:
 rule125_bdy: A125 > Agreec(A125)
 rule125_hbdy: B125 & C125
 rule125_rbdy: D125 & E125
 rule126_def:
 rule126_bdy: A126 > Agreec(A126)
 rule126_hbdy: B126 & C126
 rule126_rbdy: D126 & E126
 rule127_def:
 rule127_bdy: A127 > Agreec(A127)
 rule127_hbdy: B127 & C127
 rule127_rbdy: D127 & E127
 rule128_def:
 rule128_bdy: A128 > Agreec(A128)
 rule128_hbdy: B128 & C128
 rule128_rbdy: D128 & E128
 rule129_def:
 rule129_bdy: A129 > Agreec(A129)
 rule129_hbdy: B129 & C129
 rule129_rbdy: D129 & E129
 rule130_def:
 rule130_bdy: A130 > Agreec(A130)
 rule130_hbdy: B130 & C130
 rule130_rbdy: D130 & E130
 rule131_def:
 rule131_bdy: A131 > Agreec(A131)
 rule131_hbdy: B131 & C131
 rule131_rbdy: D131 & E131
 rule132_def:
 rule132_bdy: A132 > Agreec(A132)
 rule132_hbdy: B132 & C132
 rule132_rbdy: D132 & E132
 rule133_def:
 rule133_bdy: A133 > Agreec(A133)
 rule133_hbdy: B133 & C133
 rule133_rbdy: D133 & E133
 rule134_def:
 rule134_bdy: A134 > Agreec(A134)
 rule134_hbdy: B134 & C134
 rule134_rbdy: D134 & E134
 rule135_def:
 rule135_bdy: A135 > Agreec(A135)
 rule135_hbdy: B135 & C135
 rule135_rbdy: D135 & E135
 rule136_def:
 rule136_bdy: A136 > Agreec(A136)
 rule136_hbdy: B136 & C136
 rule136_rbdy: D136 & E136
 rule137_def:
 rule137_bdy: A137 > Agreec(A137)
 rule137_hbdy: B137 & C137
 rule137_rbdy: D137 & E137
 rule138_def:
 rule138_bdy: A138 > Agreec(A138)
 rule138_hbdy: B138 & C138
 rule138_rbdy: D138 & E138
 rule139_def:
 rule139_bdy: A139 > Agreec(A139)
 rule139_hbdy: B139 & C139
 rule139_rbdy: D139 & E139
 rule140_def:
 rule140_bdy: A140 > Agreec(A140)
 rule140_hbdy: B140 & C140
 rule140_rbdy: D140 & E140
 rule141_def:
 rule141_bdy: A141 > Agreec(A141)
 rule141_hbdy: B141 & C141
 rule141_rbdy: D141 & E141
 rule142_def:
 rule142_bdy: A142 > Agreec(A142)
 rule142_hbdy: B142 & C142
 rule142_rbdy: D142 & E142
 rule143_def:
 rule143_bdy: A143 > Agreec(A143)
 rule143_hbdy: B143 & C143
 rule143_rbdy: D143 & E143
 rule144_def:
 rule144_bdy: A144 > Agreec(A144)
 rule144_hbdy: B144 & C144
 rule144_rbdy: D144 & E144
 rule145_def:
 rule145_bdy: A145 > Agreec(A145)
 rule145_hbdy: B145 & C145
 rule145_rbdy: D145 & E145
 rule146_def:
 rule146_bdy: A146 > Agreec(A146)
 rule146_hbdy: B146 & C146
 rule146_rbdy: D146 & E146
 rule147_def:
 rule147_bdy: A147 > Agreec(A147)
 rule147_hbdy: B147 & C147
 rule147_rbdy: D147 & E147
 rule148_def:
 rule148_bdy: A148 > Agreec(A148)
 rule148_hbdy: B148 & C148
 rule148_rbdy: D148 & E148
 rule149_def:
 rule149_bdy: A149 > Agreec(A149)
 rule149_hbdy: B149 & C149
 rule149_rbdy: D149 & E149
 rule150_def:
 rule150_bdy: A150 > Agreec(A150)
 rule150_hbdy: B150 & C150
 rule150_rbdy: D150 & E150
 rule151_def:
 rule151_bdy: A151 > Agreec(A151)
 rule151_hbdy: B151 & C151
 rule151_rbdy: D151 & E151
 rule152_def:
 rule152_bdy: A152 > Agreec(A152)
 rule152_hbdy: B152 & C152
 rule152_rbdy: D152 & E152
 rule153_def:
 rule153_bdy: A153 > Agreec(A153)
 rule153_hbdy: B153 & C153
 rule153_rbdy: D153 & E153
 rule154_def:
 rule154_bdy: A154 > Agreec(A154)
 rule154_hbdy: B154 & C154
 rule154_rbdy: D154 & E154
 rule155_def:
 rule155_bdy: A155 > Agreec(A155)
 rule155_hbdy: B155 & C155
 rule155_rbdy: D155 & E155
 rule156_def:
 rule156_bdy: A156 > Agreec(A156)
 rule156_hbdy: B156 & C156
 rule156_rbdy: D156 & E156
 rule157_def:
 rule157_bdy: A157 > Agreec(A157)
 rule157_hbdy: B157 & C157
 rule157_rbdy: D157 & E157
 rule158_def:
 rule158_bdy: A158 > Agreec(A158)
 rule158_hbdy: B158 & C158
 rule158_rbdy: D158 & E158
 rule159_def:
 rule159_bdy: A159 > Agreec(A159)
 rule159_hbdy: B159 & C159
 rule159_rbdy: D159 & E159
 rule160_def:
 rule160_bdy: A160 > Agreec(A160)
 rule160_hbdy: B160 & C160
 rule160_rbdy: D160 & E160
 rule161_def:
 rule161_bdy: A161 > Agreec(A161)
 rule161_hbdy: B161 & C161
 rule161_rbdy: D161 & E161
 rule162_def:
 rule162_bdy: A162 > Agreec(A162)
 rule162_hbdy: B162 & C162
 rule162_rbdy: D162 & E162
 rule163_def:
 rule163_bdy: A163 > Agreec(A163)
 rule163_hbdy: B163 & C163
 rule163_rbdy: D163 & E163
 rule164_def:
 rule164_bdy: A164 > Agreec(A164)
 rule164_hbdy: B164 & C164
 rule164_rbdy: D164 & E164
 rule165_def:
 rule1
```

## Revisiting Simple Neural Probabilistic Language Models

Simeng Sun, Mohit Iyyer

[https://github.com/SimengSun/  
revisit-nplm](https://github.com/SimengSun/revisit-nplm)
**KEY TAKEAWAYS**

NPLM (Bengio et al., 2003) is better than expected with

- Increased depth and dimensions
- better optimization
- larger local window
- parameter reduction
- global representations

**Two Transformer-based variants**

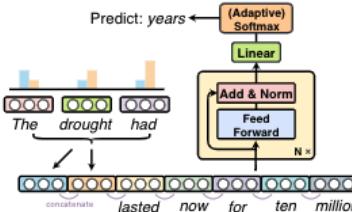
- Both variants mimic concatenation of local embeddings
- Both variants perform better than the Transformer on three word-level benchmarks.

**ABALATION**

| Model Config               | # params | Val. Perplexity |
|----------------------------|----------|-----------------|
| Transformer                | 148M     | <b>25.0</b>     |
| NPLM-old                   | 32M      | 216.0           |
| NPLM-old (large)           | 221M     | 128.2           |
| NPLM 1L                    | 123M     | 52.8            |
| NPLM 4L                    | 128M     | 38.3            |
| NPLM 16L                   | 148M     | <b>31.7</b>     |
| - Residual connection      | 148M     | 660.0           |
| - Adam, + SGD              | 148M     | 418.5           |
| - Global embedding         | 146M     | 41.9            |
| - Global kernel, + Average | 148M     | 37.7            |
| - Layer norm               | 148M     | 33.0            |

Proper optimization and residual connections are crucial to deep NPLM.

Concatenating global representations is helpful but limited compared to the Transformer

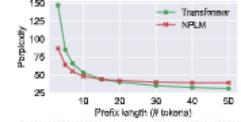
**MODERNIZED NPLM****MAIN RESULTS**

- Three word-level benchmarks (WIKITEXT-2, WIKITEXT-103, LAMBADA) and one character-level benchmark (ENWIKI).

| Model Config  | WIKITEXT-2   | WIKITEXT-103 | LAMBADA     | ENWIKI      |
|---------------|--------------|--------------|-------------|-------------|
| NPLM          | 120.5        | 31.7         | 44.8        | 1.63        |
| Transformer   | 117.6        | 25.0         | 42.1        | <b>1.14</b> |
| Transformer-C | 113.1        | 24.1         | 42.0        | 1.14        |
| Transformer-N | <b>110.8</b> | <b>24.1</b>  | <b>41.8</b> | 1.14        |

**LESSONS LEARNED**

- Old models are not that bad if scaled with modern techniques.
- Transformer variants perform better on word-level benchmarks.
- There is still a significant gap between NPLM and Transformer.
- NPLM is incapable of handling long-term context.

**TWO TRANSFORMER VARIANTS**

- NPLM achieves better perplexity when the prefix length is small
- Two variants inspired by this observation

**Transformer-N:** The first layer is the concatenation layer of NPLM  
**Transformer-C:** Local attention mask is applied to only the first layer  
 The rest of both variants are standard Transformer layers.

**ANALYSIS ON LAMBADA**

| Model         | Test F1      | Control F1   |
|---------------|--------------|--------------|
| NPLM          | 0.4          | 30.46        |
| Transformer   | 30.60        | 35.84        |
| Transformer-N | <b>32.51</b> | 37.06        |
| Transformer-C | 32.23        | <b>37.34</b> |
| Token Type    | CF F1        | LF F1        |
| Transformer   | 38.94        | 29.47        |
| Transformer-N | 42.33        | 30.14        |
| Transformer-C | <b>42.65</b> | <b>31.58</b> |
| Ent. F1       |              |              |

LAMBADA (Paper et al. 2016) test set is designed to test model's ability to understand long-term contexts.  
 We find both Transformer variants perform better for context-frequent (CF), low-frequency (LF), and named entity (Ent.) tokens.

## Limitations of autoregressive models and their alternatives

### Limitations of Autoregressive Models and Their Alternatives

Chu-Cheng Lin\*, Aaron Jaech\*, Xin Li\*, Matt Gormley , Jason Eisner\*

\*UCLingers.cs.jhu.edu

<sup>†</sup>Johns Hopkins University  
<sup>‡</sup>Facebook AI  
 Carnegie Mellon University

**Commonly held beliefs:**

"RNN language models are Turing-complete. So they can model any computable language!"  
 "RNNs can fit any finite language. If they do not fit, just add more parameters!"

**This work:**

Not really! Even with unlimited compute/annotation during training, there is a distribution over strings, that cannot be fit by any autoregressive model (e.g., RNN/Transformer), even if you allow longer strings to use larger models (with polynomial growth).  
 But this language can be easily "fit" by a short hand-written Python program!

**P:**

a decision problem class. It is the set of all languages that can be decided in polynomial time.

**Efficiently Computable (EC):**

an abstraction of Energy-Based Models (EBMs). A normalizable efficiently computable weighted language defines  $p(x) \propto f(x)$  where  $f(x)$  can be computed in  $O(\text{poly}(|x|))$ . Their support can be (and can only be) anything in P.

**Efficiently Locally Normalized (ELN):**

an abstraction of Autoregressive Models (including ordinary RNNs/LSTMs/Transformers,...) They parametrize probability of string  $x$  as  $p(x) = \prod_i p(x_i | x_{<i})$  with a fixed size parameter vector. Computing  $p(x_i | x_{<i})$  takes  $O(\text{poly}(t))$ .

**Efficiently Locally Normalizable with Compact Parameters (ELNCP):**

a generalization of ELNs. An ELNCP model has infinitely many parameter vectors. When  $|x| = n$ , an ELNCP model uses parameters  $\theta_n$  to compute  $p(x) = \prod_i p(x_i | x_{<i}, \theta_n)$ . They provide a conceptual upper bound to the just-train-a-slightly-larger-model paradigm of EBMs for autoregressive models.

ELNCP weighted languages can have support outside of P because of the precomputed parameters.

**Why is it bad that ELNCP models can't decide all languages in P?**

Because then they can't choose among continuations of a prompt. That is, there's no way to ensure that  $p(x|y) > 0$  iff  $y$  is a valid continuation of prompt  $x$ , even if that property can be checked in polytime.

**P/poly:**

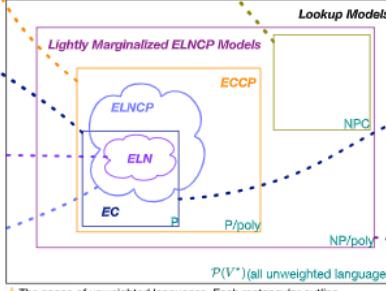
P with the help of poly-sized advice strings that can come from an oracle. P/poly is therefore more powerful than P – they can model undecidable problems due to the oracle access!

**Efficiently Computable with Compact Parameters (ECCP):**

is a generalization of ECs. Similar to ELNCPs, ECCPs is a conceptual upper bound to the just-train-a-slightly-larger-model paradigm of EBMs.

**NP-complete (NPC):**

is a set of languages that are widely believed to be outside P/poly (and therefore cannot be supported by ECCP languages)



The space of unweighted languages. Each rectangular outline corresponds to a complexity class and encloses the languages whose decision problems fall into that class. Each shape (whose name is colored to match the shape outline) corresponds to a model family and encloses the languages that can be expressed as the support of some weighted language in that family.

**Future work:**

Average-case analysis?  
 Are there model families that have all the good stuff but none of the bad stuff?

**The sequence order problem:**

consider the distribution  $p(\underline{x}, \# | \underline{y}, \#)$   
 problem encoding solution encoding

where  $p(\underline{x}, \# | \underline{y}, \#) \neq 0$  iff  $\underline{y}$  is a solution to  $\underline{x}$ .

The prefix probability  $p(x\#) > 0$  if and only if  $x$  has a solution. In other words, autoregressive models that factor  $p(x\#y) = p(x\#) \cdot p(y | x\#)$  must have the capacity to decide whether  $x$  has a solution, to ensure the joint distribution is accurate. If  $x$  is hard enough (e.g. NP-hard), no autoregressive models can even get the support right, as long as they use polytime/polylog (i.e. ELN/ELNCPs)! The other sequence order does fine under autoregressive models (if  $x$  is in NP):

$p(\underline{y}, \# | \underline{x}, \#) = p(y\#) \cdot p(x | y\#)$   
 solution encoding problem encoding

But we don't always get to decide the sequence order 😊

**Fix #1: use EBMs**

EBMs do not suffer the sequence order problem because they don't even try to compute the possibly expensive factors  $p(x_i | x_{<i})$ . Downside: It is not easy to sample from EBMs. Training them requires estimating the partition function.

**Fix #2: marginalize**

A Lightly marginalized ELNCP model marginalizes over an ELNCP language (*lightly* so because it does not have too many latent variables). The sequence of latent and observed symbols can be sampled from the ELNCP model.

Intuitively, they avoid the sequence order problem with latent variables:

$p(\underline{x}, \# | \underline{y}, \#) = p(\underline{y}, \#) \cdot p(\underline{x} | \underline{y}, \#)$   
 problem encoding solution encoding

They can have any language in NP/poly as support!

$p(\underline{x}, \# | \underline{y}, \#) = \sum_{v \in V^*} p(\underline{y}, \#) \cdot p(\underline{x} | \underline{y}, \#)$   
 problem encoding solution encoding

Downside: marginalization is required even at test time.

**Fix #3: memorize anything we need**

We can model anything if we have a big big database!  
 Examples: kNLM, adaptive semi-parametric language models....  
 Downside: Need a vast database of observed or precomputed answers.

Model family      Compute parameters      Efficient sampling      Efficient sampling and memorization      Support size ...

ELNCP: Autoregressive models (P) ✓      ✓      ✓      ✓      never but not all L & P

ELNCP: Energy-based models (NP) ✓      ✓      ✓      ✓      all L & P but not L & NP

Lightly marginalized ELNCP: Latent variable autoregressive models (ELNCP) ✓      ✓      ✓      ✓      all L & NP

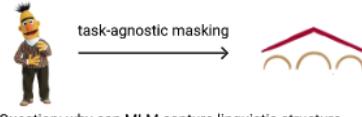
Lightly marginalized ELNCP: Latent variable energy-based models (ELNCP) ✓      ✓      ✓      ✓      everything

# On the inductive bias of masked language modeling: from statistical to syntactic dependencies

## On the Inductive Bias of Masked Language Modeling: From Statistical to Syntactic Dependencies

Tianyi Zhang, Tatsunori B. Hashimoto

### 1 Problem Statement



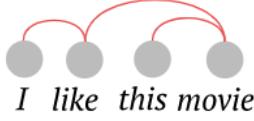
Question: why can MLM capture linguistic structure and transfer to new tasks?

### 2.1 Cloze Reduction Hypothesis



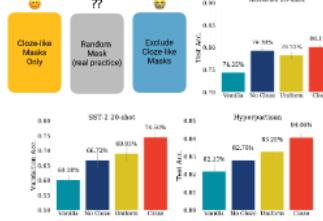
Cloze-like masking can provide indirect supervision but in practice we apply random masking.  
We quantify the importance of cloze-like masking through controlled experiments.

### 2.2 Dependency Learning Hypothesis



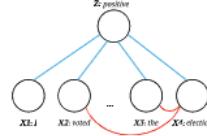
We hypothesize that generic masks can help learn the statistical dependencies among words and these dependencies are related to syntactic structures

### 3 Uniform vs. Cloze-like Masking



A substantial part of performance gain comes from generic masking. The cloze reduction hypothesis alone cannot account for the entire success of MLM

### 4 MLM as Dependency Learning



### 4.1 MLM recovers latent variables

MLM representations are similar to representations obtained by supervised learning (with access to Z)

**Proposition 1.** Assuming that  $\Sigma_{XX}$  is full rank,

$$x_{mask,i} = \beta_{2SLS,i} X_{\setminus i} + O(\|\Sigma_{XX \setminus i,i}\|_2),$$



code release@github:  
tatsu-lab/mlm\_inductive\_bias

### 4.2 MLM recovers direct dependencies

Cond. MI reveals direct dependencies in presence of latent variables.

**Proposition 3.** The gap between conditional MI with and without latent variables is bounded by the conditional entropy  $H(Z | X_{\setminus \{i,j\}})$ .

$$I(x_i; x_j | X_{\setminus \{i,j\}}) - I(x_i; x_j | Z, X_{\setminus \{i,j\}}) \leq 2H(Z | X_{\setminus \{i,j\}}).$$

MLM objective directly ensures good approximation of cond. MI.

**Proposition 4.** Let

$$\hat{I}_{pq} = \mathbb{E}_{x_i, x_j} [\log p_{\theta}(x_i | X_{\setminus \{i,j\}}) - \log \mathbb{E}_{x_j | x_i} p_{\theta}(x_i | X_{\setminus \{j,x_i\}})]$$

be an estimator constructed by the model distribution  $p_{\theta}$ . Then we can show,

$$|\hat{I}_{pq} - I_{pq}| \leq \mathbb{E}_{x_i} [p(x_i | X_{\setminus \{j,x_i\}}) \| p_{\theta}(x_i | X_{\setminus \{j,x_i\}}) ].$$

### 5 Statistical Dependencies are related to Syntactic Dependencies

We extract the cond. MI from a pretrained BERT model.

We convert cond. MI to unsupervised parses and show that the resulting trees are related to dependency parses.

| Method                   | UUAS                               |
|--------------------------|------------------------------------|
| RANDOM                   | $28.50 \pm 0.73$                   |
| LINEARCHAIN              | 54.13                              |
| Klein and Manning (2004) | $55.91 \pm 0.68$                   |
| PMI                      | 33.94                              |
| CONDITIONAL PMI          | $52.44 \pm 0.19$                   |
| CONDITIONAL MI           | <b><math>58.74 \pm 0.22</math></b> |

Table 1: Unlabeled Undirected Attachment Score on WSJ10 test split (section 2.3). Error bars show standard deviation across three random seeds.