

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 semantic 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

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

<h3>Introduction</h3> <ul style="list-style-type: none"> Explicit abuse is abuse conveyed by abusive words: <ul style="list-style-type: none"> Stop editing this, you dumbass. Go lick a pig, you arab piece of scum. Implicit abuse is abuse not conveyed by abusive words: <ul style="list-style-type: none"> You run like a headless chicken. Black people steal everything. State-of-the-art classifiers are good at detecting explicit abuse. There is no indication that classifiers are able to detect implicit abuse reliably (Wiegand et al., 2019). 	<h3>Different Subtypes of Implicit Abuse</h3>	<h3>Identity-Term Bias (I)</h3> <ul style="list-style-type: none"> Datasets are often created by sampling for topics that are likely to contain abusive language. This sampling distorts the word distribution and causes spurious correlations. <p>Thus, identity terms often only occur in the abusive data.</p>	<h3>Identity-Term Bias (II)</h3> <p>Statistics computed on the Social Bias Frame Corpus (Sap et al., 2020):</p>
<h3>Joke Bias (Social Bias Frame Corpus)</h3> <ul style="list-style-type: none"> Jokes mostly originate from subreddits on abusive jokes. Jokes have a common structure on that dataset: <ul style="list-style-type: none"> What's worse than an angry black woman? Nothing. What's better than winning the Paralympics? Walking. Structure: question-word word word ... ? word. We only observe this pattern on abusive microposts. Supervised classifiers will learn this pattern as implicit abuse. Supervised classifiers are unlikely to learn to understand jokes or to detect implicit abuse. 	<h3>What happens if we merge different datasets?</h3> <ul style="list-style-type: none"> Sometimes, data from different sources are merged in order to obtain a larger dataset. Each data source has its unique style, word distribution and class distribution. Merging different datasets may result in classifiers learning to detect the individual data sources. In an example on the left: <ul style="list-style-type: none"> dataset₁ is a proxy for abuse dataset₂ is a proxy for non-abuse 	<h3>More training data does not necessarily mean better data!</h3> <ul style="list-style-type: none"> A smaller unbiased dataset is better than a large biased dataset! No use in expanding dataset if only frequently occurring phenomena are repeatedly added. 	<h3>Divide and Conquer</h3> <ul style="list-style-type: none"> Create one dataset for each subtype of implicit abuse. Train individual classifiers on these datasets and then merge the predictions. Advantage: less heterogeneous set of abusive data should make it easier to produce appropriate negative data. Thus there is a better chance that the resulting datasets are less biased.
<h3>Filter Out Explicit Abuse</h3> <ul style="list-style-type: none"> Often implicit abuse co-occurs with explicit abuse: <ul style="list-style-type: none"> Sneaky [Jews control the world] Supervised classifiers will learn to detect the explicit clues from those instances because they are much easier to spot. The implicit clues will be ignored. Suggestion: manually remove the explicit clues from the classifiers and train on these edited data. This is the only chance that supervised classifiers will consider learning the implicit clues. 	<h3>Transform Mentions into Uses</h3> <ul style="list-style-type: none"> Large proportion of implicit abuse contained in non-abusive microposts, because they are reported cases. Classifiers will never have a chance to learn these forms of implicit abuse from those instances. Classifiers will simply learn clues indicating reported speech as a signal for non-abusive language. Suggestion: manually isolate the implicitly abusive utterances from those datasets and use them as additional abusive training data. 	<h3>Conclusion</h3> <ul style="list-style-type: none"> Many different subtypes of implicit abuse. Either too infrequent in available datasets or biases prevent classifiers from truly learning them. Merging existing datasets does not help. Large datasets are not the solution to the problem either. We need novel datasets, ideally one for each subtype and properly debiased. For new datasets negative data should be carefully sampled. 	

The importance of modeling social factors of language

erc
Dirk Hovy
Bocconi University
Milan
Italy

The Importance of Modeling Social Factors of Language: Theory and Practice

Diyi Yang
Georgia Tech
Atlanta, GA
USA

HYPOTHESIS: CURRENT NLP APPROACHES IGNORE THE SOCIAL FACTORS OF LANGUAGE, THAT LIMITS PERFORMANCE AND POSSIBLE APPLICATIONS.

Social Norms

Acceptable conduct, shared understanding.

Communicative Goals

Metafunction of ideational and interpersonal goals.

Culture & Ideology

Customs, ideology, and cultural identity.

CORRECT, BUT UNPOLITE
I don't know?
Where's the pharmacy?

INCORRECT, BUT POLITE
Take the first left, walk down, go right at the intersection, and keep on going if you see a large tree, then...

Social Relation

Distance and relation nature between speakers.

communicative goals
 culture & ideology
 social norms
 context
 social relation
 speaker receiver

Context

Non-textual factors: domain, language, situation, topic, etc.

Speaker

Consistent traits signaling demographic makeup. Improve NLP performance

Receiver

Intended audience and their expectations.

Hey, can't make tonite, sorry!

VS

Dear Madam President, I regret to inform you that I will not be able to participate.

Bocconi

Preregistering NLP research

Preregistering NLP Research

Emiel van Miltenburg, Chris van der Lee, Emiel Kraemer

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
- ✗ 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.
 - Checks 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:

The keys to the cabinet _____ on the table.

	P_M
are	0.6
is	0.05
exist	0.1
exists	0.25

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

$P(\text{are}) > P(\text{is})$

$P(\text{exist}) < P(\text{exists})$ ← TSE misses this because it's not in the minimal pair dataset.

Our Metrics

Consider:

- a minimal pair context c
- a set of Lemmas \mathcal{L} where $\ell \in \mathcal{L}$
- ℓ_+ is the correct conjugation and
- a model P_M
- ℓ_- is the incorrect conjugation

Likely Behavior: Model Weighted Syntactic Evaluation (MW)

$$MW = \frac{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ | c)}{\sum_{\ell \in \mathcal{L}} P_M(\ell_+ | c) + P_M(\ell_- | c)}$$

E.g. $\frac{P_M(\text{are}) + P_M(\text{exist})}{P_M(\text{are}) + P_M(\text{is}) + P_M(\text{exist}) + P_M(\text{exists})} = 0.7$

Systematicity: Equally Weighted Syntactic Evaluation (EW)

$$EW = \sum_{\ell \in \mathcal{L}} \frac{\mathbb{1}[P_M(\ell_+ | c) > P_M(\ell_- | c)]}{|\mathcal{L}|}$$

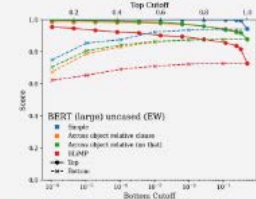
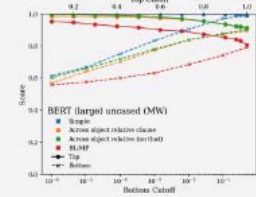
E.g. $\frac{1}{2} (\mathbb{1}[P_M(\text{are}) > P_M(\text{is})] + \mathbb{1}[P_M(\text{exist}) > P_M(\text{exists})]) = 0.5$

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%



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 (uncased)			RoBERTa		
	MW	EW	TSE	MW	EW	TSE
Simple	0.99	0.94	1.00	0.98	0.90	1.00
In a sentential complement	0.92	0.67	0.89	0.92	0.60	0.86
VP coordination	0.91	0.89	0.90	0.93	0.90	0.90
Across prepositional phrase	0.91	0.83	0.93	0.83	0.75	0.85
Across subject relative clause	0.87	0.84	0.84	0.88	0.84	0.85
Across object relative clause	0.91	0.88	0.91	0.86	0.80	0.85
Across object relative (no that)	0.92	0.88	0.90	0.79	0.72	0.81
In object relative clause	0.93	0.95	0.97	0.95	0.97	0.99
In object relative (no that)	0.90	0.91	0.92	0.81	0.82	0.82
BLIMP	0.81	0.73	0.90	0.78	0.69	0.85

Qualitative Examples

	simple	complex	non	sent	coord	sub	obj	fact	adv	adv
The reason that the doctor was not young	0.28	0.02	0.02	0.08	0.03	0.03	0.04	0.03	0.02	0.03
The photo that the executive was not tall	0.04	0.04	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03

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] The keys to the cabinet ~~was~~ not young.
[2] Marvin and Linzen, 2018. Targeted syntactic evaluation of language models. *EMNLP*.
[3] Richard Stebbins, 2009. *BLIMP: A Benchmark for English Grammatical Pairs for English NLP*.
[4] The use of word or phrase in the context of the sentence. *COCA*.
[5] The Penn Treebank. 2009. The syntax of contemporary written English.
[6] Michael R. Beaulieu, Richard Stebbins, and Mary Ann Moens. 1985. Building a large annotated corpus of English. *The Penn Treebank*, 31.
[7] John Hewitt, John Wang, Clark & Emily Healy. 2019. Good verbs list: 3,200 verbs plus spelling, class, and singular verbs marked.

Adaptable and interpretable neural memory over symbolic knowledge

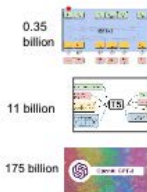
Adaptable and Interpretable Neural Memory Over Symbolic Knowledge

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



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



Model

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)

Dataset	Full Dataset		WikiData Answerable	
	Total	No Overlap	Total	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	28.6	-

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 on updated facts	+ updated memory
	0.0	54.5

Q: Where was [Charles Darwin] born?
 A: [United-Kingdom]

Q: Where was [Charles Darwin] born?
 A: [Germany]

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



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.

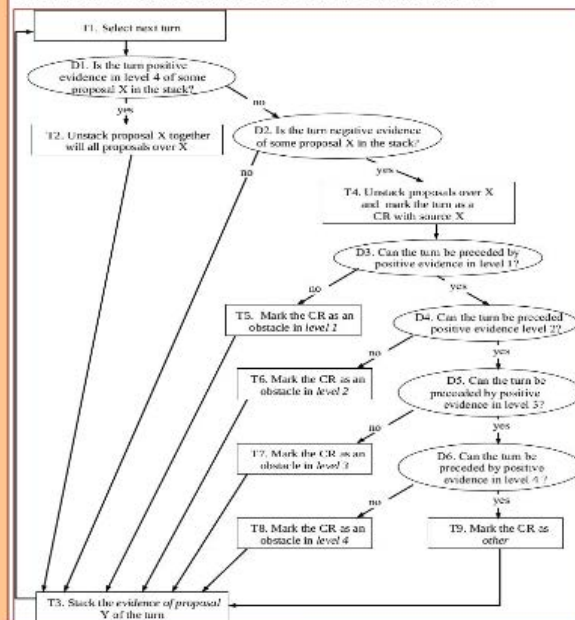


2. Clarification mechanisms make this interpretative process explicit - they ground utterances in world modalities such as vision, touch, movement,...

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 .



Causal effects of linguistic properties

Causal Effects of Linguistic Properties

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, Dhanya Sridhar
Stanford University, University of Chicago, Columbia University

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:

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

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 sentiment 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.

Estimation

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

- Use a form of distant supervision to improve the quality of the proxy labels \hat{T} .
- Train a neural network to predict both of Y 's potential outcomes when \hat{T} is 0 and 1.
- Run inference over a test set and calculate the average difference between the two potential outcomes.

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].

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. This error is a positive term meaning our estimate only attenuates the true ATE:

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

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].

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 complaint politeness on response time?"

References

- Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- Zach Wood-Doughty, Ilya Shtatman, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of EMNLP*.
- Nasir Egeci, Christian Fong, Justin Gimenez-Margaret, E Roberts, and Brandon H Stewart. 2018. How to make causal inferences using texts. arXiv preprint arXiv:1802.02143.

rpryzant@stanford.edu

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

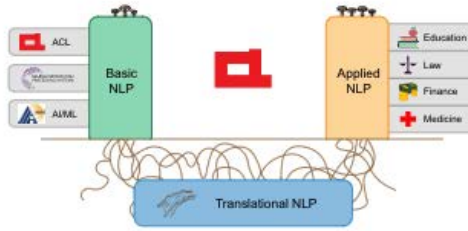
Translational NLP: A New Paradigm and General Principles for Natural Language Processing Research

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



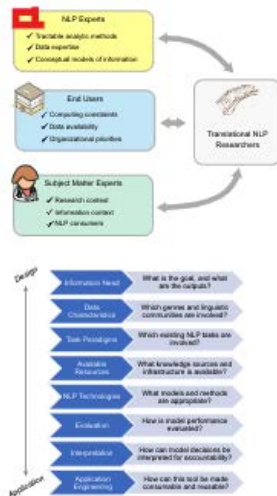
Abstract

- 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

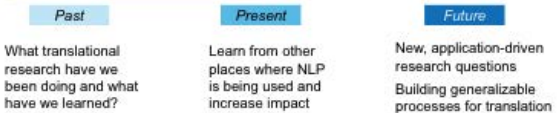
Our questions are a **starting point** for evolving translational NLP development.

What does it look like?

Translational NLP projects have three components:



What's next?



Papers in interpretability

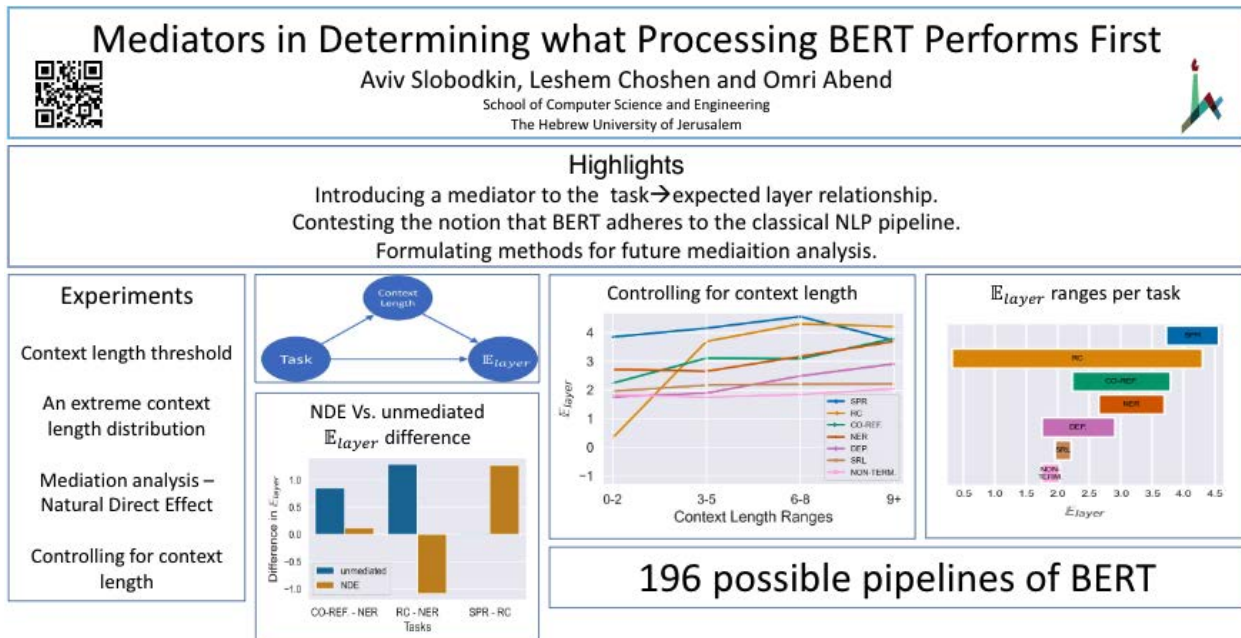
Session 1B

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

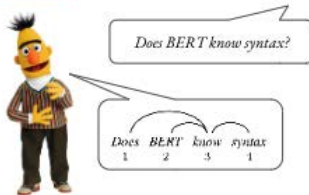


Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing

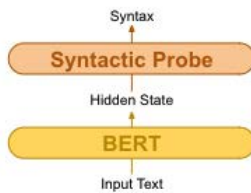


Rowan Hall Maudslay Ryan Cotterell

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



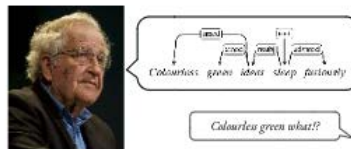
[2/7] It does this by training a supervised model to predict syntax using another model's hidden state; if it can do this, then people argue those representations contain syntax.



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



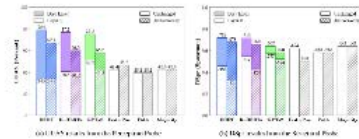
[4/7] The trouble is, the sentences used for probing are real-world sentences, which do not properly isolate syntax. Chomsky:



[5/7] To investigate whether probes leverage semantic patterns to aid 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>

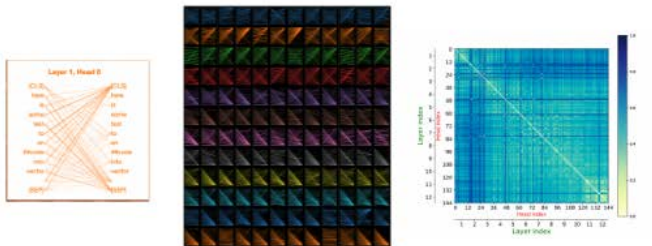
- Task: Given a configuration, move a block from the source location to the target location.
- Aim to understand if the test performance of these models indicates an understanding of the spatial domain and of the natural language instructions relative to it, or whether they merely overfit spurious signals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

On attention redundancy: a comprehensive study

On Attention Redundancy: A Comprehensive Study NAACL 2021

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

Motivation



Attention head

L-layer-H-head

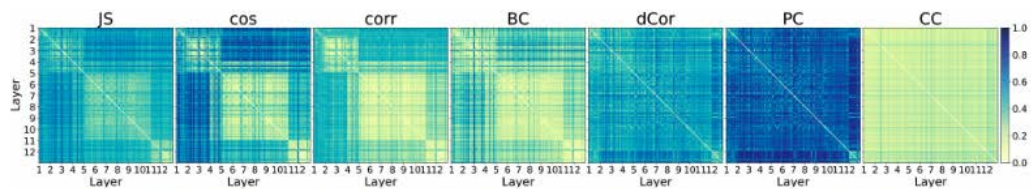
Redundancy matrix (What)

Method: 5-Ws and How

- **What** is attention redundancy?
 - Redundancy matrix
- **How** to measure attention redundancy?
 - Distance functions: sentence-based, token-based
- **Where** does attention redundancy exist?
 - Cluster patterns
- **When** does attention redundancy occur?
 - Phase-independent: Pre-trained vs Fine-tuned
- **Who** (which task) has attention redundancy?
 - Downstream task agnostic
 - Case study: zero-shot head-pruning strategy
- **“Why”** does attention redundancy happen?
 - Influences of dropout ratios during pretraining

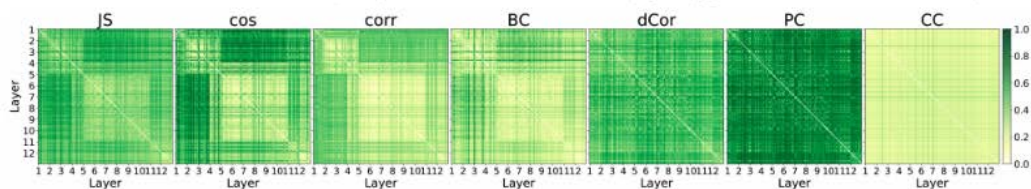
Experimental Results

How & Where



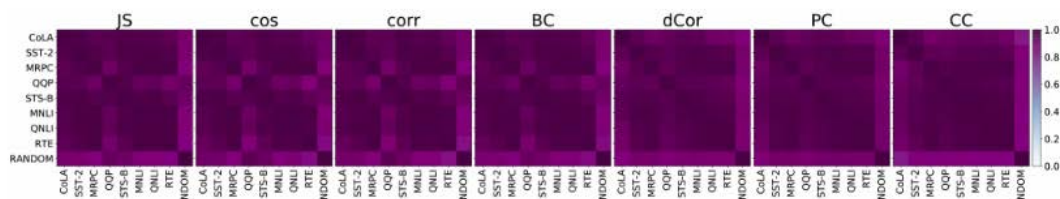
How and Where: token-based (left 4) and sentence-based distances (right 3) (pre-trained BERT-base on CoLA)

When



When: Phase-independent: Fine-tuned (fine-tuned BERT-base on CoLA) v.s. Pre-trained (above)

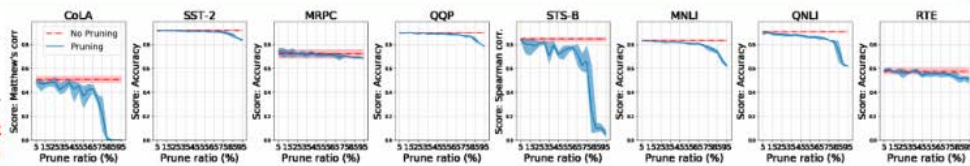
Who



Who: Task-agnostic: Correlations of redundancy matrices of each pair of tasks (pre-trained BERT-base on GLUE tasks)

Zero-shot head pruning:

- **Input:**
 - Redundancy matrices
 - Model: pretrained BERT
 - Data: randomly generated token sequences (no data of downstream tasks is required)



Fine-tuning results of pruned BERT-base model on GLUE tasks
(75 – 85% heads can be pruned to preserve performances for most tasks)

Steps:

- Clustering based on red redundancy matrices
- Get cluster center object (left attn head) based on the given cluster goodness metric

Chances of a head being pruned at various pruning ratios (averaged over 10 clustering runs)

“Why”: dropout ratios?

- Token-based distance (JS)
 - “N”-shape
- Sentence-based distance (dCor)
 - Monotonic effects
- Sensitivity
 - hidden-dropout > attention-head-dropout

References

- Olga Kovalova, Alexay Romanov, Anna Rogers, and Anna Rumshisky. 2019. *Revealing the dark secrets of bert*. In EMNLP
- Paul Michol, Omor Levy, and Graham Neubig. 2019. *Are sixteen heads really better than one?* In NeurIPS.
- Jacob Devlin, Ming-Wai Chang, Kenton Loo, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In NAACL.
- Kevin Clark, Urvashi Khandelwal, Omor Levy, and Christopher D Manning. 2019. *What does bert look at? an analysis of bert’s attention*. arXiv preprint arXiv:1906.04341

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, rdeshpan, franck.dernoncourt, jigu, tsun}@adobe.com



Shortcut Learning In BERT-based NLU Tasks

- The NLU (natural language understanding) Task

Premise: Trevor Griffiths was born on April 4, 1935

Hypothesis: Trevor Griffiths (born 4 April 1935), is an English dramatist.

⚡ Contradiction
⚡ Entailment
⚡ Neutral
- The Explanation for BERT-based NLU models

neutral (1.00)	[CLS] no not near as much as i'd like to i mean i've i tend to stay pretty busy at my job and uh [SEP] my job wasn't too busy, i do that a lot more. [SEP]
entailment (0.67)	[CLS] equivalent to increasing national saving to 19. [SEP] national savings are equivalent . [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] the record gave no context regarding the legal question. [SEP]
contradiction (0.68)	[CLS] hellenic and roman periods [SEP] the hellenic period. [SEP]
neutral (0.89)	[CLS] he thought phil hundert thing husbands 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 the uh education system i forget 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]
- 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

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

Long-tailed distribution

Data statistics

Model behavior

Shortcut degree

Example of model paying high attention to features on the head

[CLS] the lot upon which it is being built had been vacant. [SEP]

[SEP] the lot had been vacant. [SEP]
- Self knowledge Distillation for Mitigating Shortcut Learning
 - Dis-encourage Model to Giving Overconfident Prediction for Shortcut Samples

Teacher model

Softmax

Overconfident prediction

Student model

Softmax

Smoothed Softmax

Distill loss

Student loss

Ground truth y

Low-complexity probing via finding sub-networks

NAACL 2021 papers

16

Low-Complexity Probing via Finding Subnetworks

Steven Cao, Victor Sanh, and Alexander M. Rush



Motivation: faithful probing

- **Goal of probing:** figure out which properties are captured in pre-trained models.
- **Status quo:** freeze BERT parameters and train a shallow MLP on top to minimize task loss.
- **Idea:** High probe accuracy → property is captured.
- **Problem:** Accuracy does not always reflect whether the property is captured!
A shallow probe is unable to find many properties, while a deep probe is often able to learn on its own.
- **Key idea:** rather than **adding** parameters, we should **remove** parameters to find task subnetworks in BERT. Accurate subnetwork exists → property is captured.

Evaluation: plotting the accuracy-complexity tradeoff

- A good probe should have **high accuracy** (able to find properties in models) and **low complexity** (unable to learn on its own).
- We vary the complexity of each probe and show the resulting **complexity-accuracy plot** [Pimentel+ 2020].
- **Baseline:** MLP with 1 hidden layer.
- Vary complexity by restricting hidden layer rank.
- **Ours:** subnetwork probing.
- Vary complexity by grouping weights with a one mask.
- Complexity measured in **bits needed to transmit the probe parameters** [Voita+ 2020].
 - 1 bit per mask, between 1 and 32 bits per float

Evaluation: probing accuracy on pre-trained and random networks

Recall: Want **high pre-trained** (finds properties) and **low random** (cannot learn on its own).

Subnetwork probing: methods

$M_i = \text{HardConcrete}(\theta_i)$

$\min_{\theta, \text{CLS}} E_M [\text{Loss}(\text{BERT}_{\theta+\text{CLS}}) + \lambda \text{Sparsity}(M)]$

- Associate each neuron with mask M
- Learn M using continuous relaxation θ
- Minimize task loss while encouraging sparsity

Analysis: subnetwork location

- Plot the percentage of each layer that is un-pruned.
- **Result:** lower-level tasks are captured in lower layers, reproducing NLP pipeline result in [Tenney+ 2019]: POS → parsing → NER
- Many possibilities for further analysis!

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).

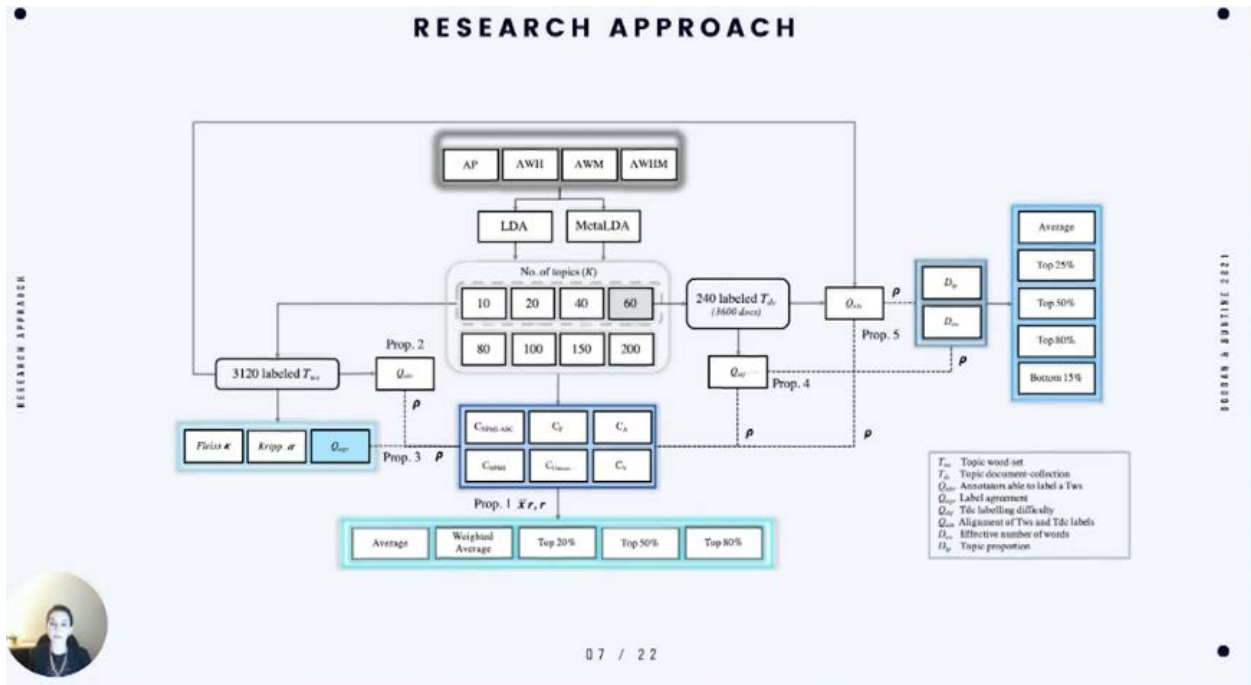
Interpretability analysis for NER to understand system predictions and how they can improve

- 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.
 - Qaln: Rated alignment between Tws and Tdc.



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

Hanjie Chen,¹ Song Feng,² Jatin Ganhotra,² Hui Wan,² Chulaka Gunasekara,² Sachindra Joshi,² Yangfeng Ji¹
 hc9mx@virginia.edu ¹Department of Computer Science, University of Virginia ²IBM 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.

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 $\min_{\theta} \mathcal{L}_{ce}(y, \hat{y}) - \gamma_1 (H(Z^U) + H(Z^L)) + \gamma_2 H(G)$
- 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

Decompose word mask

$$W_{i,j} = \sum_{g=1}^G \delta(z_{i,j}, g) \delta(g, j)$$

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

Weighted word attributions

$$\hat{\theta}_{i,j} = \sum_{g=1}^G \phi_{i,j}(g) \psi(g)$$

The expectation of $W_{i,j}$

Experiments

Setup

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

AOPC $AOPC = \frac{1}{|U|+1} \sum_{i=1}^{|U|} |P(V^{(i)}) - P(V^{(i+1)})|$ ✓ Higher AOPC is better

Models	Methods	e-SNLI	Quora	QQP	MRPC
DAtn	LIME	0.286	0.120	0.078	0.084
	L2X	0.269	0.128	0.079	0.089
	IBA	0.254	0.137	0.104	0.209
	GMASK	0.324	0.160	0.087	0.084
BERT	LIME	0.221	0.153	0.110	0.062
	L2X	0.105	0.119	0.134	0.083
	IBA	0.282	0.199	0.144	0.134
	GMASK	0.292	0.212	0.139	0.130

Post-hoc accuracy $Post-hoc-acc(y) = \frac{1}{|U|} \sum_{i=1}^{|U|} \mathbb{1}[y_i^{(u)} = y^{(u)}]$

Degradation Test

- GMASK achieves higher degradation score

Models	Methods	e-SNLI	Quora	QQP	MRPC
LIME	LIME	0.502	0.070	0.091	1.367
	L2X	0.366	0.042	0.036	1.779
	IBA	0.473	0.110	0.107	2.775
	GMASK	0.476	0.143	0.214	2.037
DAtn	LIME	0.403	0.176	0.208	2.798
	L2X	0.338	0.102	0.087	-0.018
	IBA	0.383	0.168	0.173	-0.083
	GMASK	0.369	0.303	0.172	0.251

Visualization of top four words

LIME: a man playing an electric guitar on stage. a man playing banjo on the floor
 L2X: a man playing an electric guitar on stage. a man playing banjo on the floor
 DAtn: a man playing an electric guitar on stage. a man playing banjo on the floor
 IMASK: a man playing an electric guitar on stage. a man playing banjo on the floor
 GMASK: 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/nlp-second-order-attack

Overview

Robustness and counterfactual bias are usually evaluated on a test dataset. However, are these evaluations robust?
 • A sentence x_0 from the test / dev set may be robust.
 • But it could become **vulnerable** after a slight perturbation (x_0').

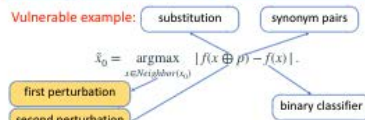


- Double perturbation**
- To uncover model weaknesses beyond the test set.
 - Perturb test set to construct similar sentences.
 - Diagnose prediction changes for a single-word substitution.
 - **Outcome:**
 - Identify vulnerable examples for 96.0%-99.8% test examples on robustly trained CNNs and Transformers.
 - Reveal the hidden model biases.

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.

Evaluating Second-Order Robustness



Step 1: Find p for x_0
 Choose p from a predefined list of counter-fitted synonyms P :
 $p = \operatorname{argmax}_{p \in P, p \neq x_0} |f(p^{(2)}) - f(p^{(1)})|$
 e.g., $x_0 =$ "a deep and meaningful film." and $p =$ (film, movie).

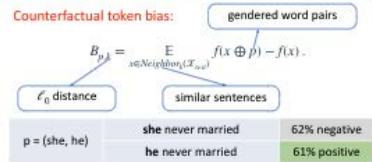
Step 2: Find Vulnerable Examples Through Beam Search
 Replace one word at a time through a masked language model.

Beam	Neighborhood Sentence	$f(x)$
1	a deep and moving film (movie).	.999 (.999)
	a dramatic and meaningful film (movie).	.999 (.999)
	a short and moving film (movie).	.730 (.303)
2	a slow and moving film (movie).	.519 (.151)
	a dramatic or meaningful film (movie).	.487 (.168)

Experimental Results

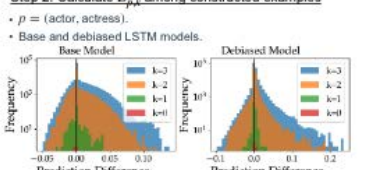


Evaluating Counterfactual Bias



Step 1: Construct abundant natural sentences
 Construct through a masked language model.

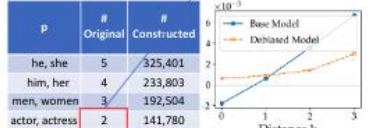
Step 2: Calculate $B_{p,x}$ among constructed examples



Reveal the Hidden Biases

- Base model:
- Negative bias when $k = 0$ (effectively the original dataset).
 - Positive bias when $k = 3$ (our method).

Why: The naive evaluation only use two test examples and thus is not robust.



- 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



On Biasing Transformer Attention Towards Monotonicity

Annette Rios¹, Chantal Amrhein¹, Noëmi Aeppli¹ and Rico Sennrich^{1,2}
¹Department of Computational Linguistics, University of Zurich
²School of Informatics, University of Edinburgh

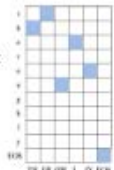


Why Monotone Attention?


- 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

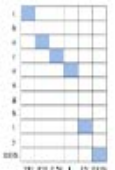
Attention α :

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{|X|} \exp(e_{ik})}$$


Mean Attended Position \bar{a}_i :

$$\bar{a}_i = \sum_{j=1}^{|X|} \alpha_{ij} \cdot j$$


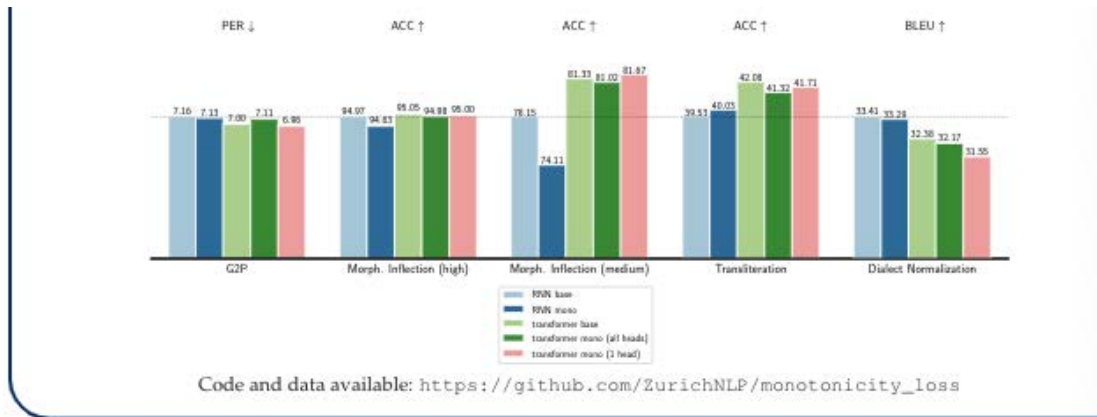
Loss:

$$L_{mono} = \sum_{i=1}^{|Y|-1} \max\left(\frac{\bar{a}_i - \bar{a}_{i+1} + \delta \frac{|X|}{|Y|}}{|X|}, 0\right)$$


δ	L_{mono}	L_{mono}	L_{mono}	L_{mono}
$\delta=0$	0.364	0.000	0.000	0.000
$\delta=0.5$	0.530	0.167	0.000	0.000
$\delta=1$	0.697	0.333	0.152	0.000

Evaluation

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



- ### 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

Finding Concept-specific Biases in Form-Meaning Associations

UNIVERSITY OF CAMBRIDGE
 Google
 UNIVERSITÄT TUBINGEN
 ETH zürich
 HARVARD UNIVERSITY

Tiago Pimentel
Brian Roark
Sören Wichmann

Ryan Cotterell
Damián Blasi

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; Phonesthemes; Iconicity.

Data - ASJP

- Basic vocabulary wordlists.
- 5189 languages! Almost ¼ 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):

$$MI(\text{meaning}, \text{form}) = H(\text{form}) - H(\text{form} | \text{meaning})$$

Cross-entropy approximations:

$$H(\text{form}) \leq H_0(\text{form}) \approx -\frac{1}{N} \sum \log p_0(\text{form})$$

$$H(\text{form} | \text{meaning}) \leq -\frac{1}{N} \sum \log p_0(\text{form} | \text{meaning})$$

Cross-linguistic Challenges

Cross-entropy needs to be computed on data independent from training! Languages have been in contact (not i.i.d.).

- We split our data per macro-area. 2 areas for training, 1 development, 1 test;
- We group language families in a single area
- We drop loan words;
- We control for language family size.

Overall results

Macroarea	H(W)	MI(W, V)	U(W V)
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%

Very small average contribution of meaning into form.
• Approximately 0.3%.

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?

How (Non-)Optimal is the Lexicon?

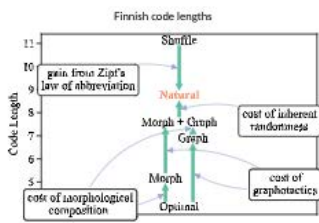
UNIVERSITY OF CAMBRIDGE
UC SANTA BARBARA
Tiago Pimentel* Irene Nikkarinen* Kyle Mahowald
Ryan Cotterell Damián E. Blasi
ETH zürich
HARVARD UNIVERSITY

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:

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

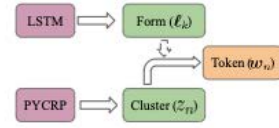
The Meaning Distribution

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:

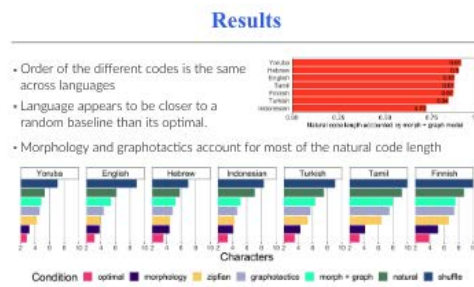
$$\text{cost} \approx \frac{1}{N} \sum_{n=1}^N |w_n| \quad \text{cost} \approx \frac{1}{N} \sum_{n=1}^N \left\lceil \log_{256} \frac{1}{p(w_n)} \right\rceil \quad \text{cost} \approx \frac{1}{N} \sum_{n=1}^N |w_n|$$

natural code
optimal code
graphotactic code

$$\text{cost} \approx \frac{1}{N} \sum_{n=1}^N \sum_{m=1}^M \left\lceil \log_{256} \frac{1}{p(w_{n,m})} \right\rceil \quad \text{cost} \approx \frac{1}{N} \sum_{n=1}^N \sum_{m=1}^M |w'_{n,m}|$$

morphology code
morph+graph code

- We rely on Morfessor to get individual "morphemes".
- We sample wordforms from an LSTM to get our graphotactics code.



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, 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.

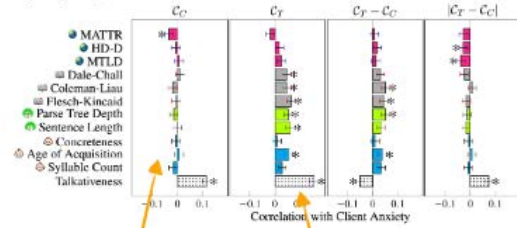
	Dataset	
	Exploratory	Confirmatory
Messages	2.6 million	0.7 million
Survey responses	24,287	6,150
Clients	5,736	1,434
Therapists	1,608	889
†Survey responses / client	4.23	4.29
†Client text (words) / survey	1259	1295
†Therapist text (words) / survey	796	804
Median survey score (0-21)	8	8
Median time between surveys	21 days	21 days

> 3 million messages

~ 4 survey responses per client indicating anxiety over time! (scored from 0-21)

Linguistic Complexity Loss

Figure 1: Linguistic complexity measures correlate with client anxiety (⊕ indicates significance at $p < 0.001$ for both the exploratory and confirmatory datasets). We show correlations ($\pm 99.9\%$ confidence intervals) on the exploratory dataset for language complexity of clients (C_C), therapists (C_T), therapist and client difference ($C_T - C_C$), and absolute therapist and client difference ($|C_T - C_C|$). Each complexity measure was entered into its own linear mixed model. We group complexity measures into lexical diversity (⊕), syntax (⊕), readability (⊕), and prosodicity (⊕).



When clients were more anxious, their linguistic complexity dropped.

When clients were more anxious, therapist linguistic complexity increased.

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_C	σ_T	σ_{C-T}	$\sigma_{ C-T }$	σ_C	σ_T	σ_{C-T}	$\sigma_{ C-T }$
⊕ MATTR	-0.36	-0.33	-0.87	0.3642	-0.35	-0.29	-2.17	0.0298
⊕ HD-D	-0.3	-0.32	0.53	0.5936	-0.3	-0.28	-0.83	0.4082
⊕ MTL	-0.36	-0.35	-0.06	0.9561	-0.35	-0.34	-0.17	0.8665
⊕ Dale-Chall	-0.46	-0.65	6.43*	<0.0001	-0.45	-0.51	1.91	0.0563
⊕ Coleman-Liau	-0.68	-0.74	1.91	0.0558	-0.66	-0.61	1.84	0.0666
⊕ Flesch-Kincaid	-0.46	-0.93	15.08*	<0.0001	-0.47	-0.76	11.53*	<0.0001
⊕ Parse Tree Depth	-0.77	-0.89	4.12*	<0.0001	-0.74	-0.73	-0.48	0.6324
⊕ Sentence Length	-0.44	-0.97	17.69*	<0.0001	-0.45	-0.79	12.56*	<0.0001
⊕ Concreteness	-0.49	-0.36	-4.84*	<0.0001	-0.48	-0.32	-5.73*	<0.0001
⊕ Age of Acquisition	-0.48	-0.69	6.15*	<0.0001	-0.47	-0.51	1.44	0.1494
⊕ Syllable Count	-0.44	-0.44	0.11	0.9913	-0.43	-0.36	-2.31	0.0373
Talkativeness	-0.31	-0.66	13.48*	<0.0001	-0.3	-0.58	11.88*	<0.0001

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

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

Data Privacy Statement

All patients are provided with a copy of their data in a de-identified, aggregate format as part of the user agreement before they begin using the platform and can opt out at any time by contacting their therapist or by contacting support. Study procedures were approved as reported by our institution's Institutional Review Board (IRB).

Transcripts were de-identified automatically via a HIPAA-compliant interface by anonymizing all proper nouns, dates, names, and other content features of language. All information related to names or content was then removed, including words, phone numbers, addresses, though these were infrequently found in the interaction between therapists and patients.

Ethical Considerations

The dataset in this paper is of a sensitive nature, and there are several associated ethical considerations. Our study procedures were approved as exempt by the Committee for the Protection of Human Subjects at our institution. All patients and therapists gave consent for the use of their data in a de-identified, aggregate format and the dataset is not publicly available. All patients were able to opt out at any time by contacting their therapist or contacting support. We emphasize that the findings in our paper are specific to the dataset and do not make any claims about their generalizability to other contexts. Our study was not intended to provide clinical or practical applications. Finally, the data that we shared were written in English and therefore we do not claim that our findings generalize to other languages. For these reasons, we advise caution when working with this dataset and building upon these results.

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

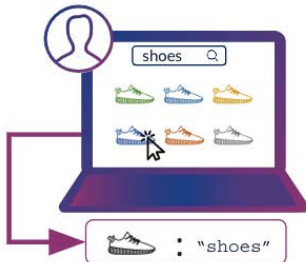
Meaning is grounded in objects

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.



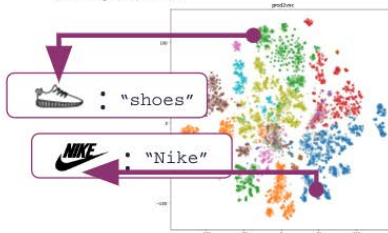
Using IR to learn a grounded semantics for noun phrases end-to-end

- 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.



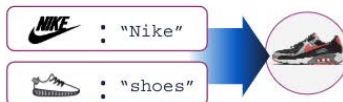
Lexical denotation as DeepSets

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 - $DeepSet \times DeepSet \rightarrow DeepSet$:



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 causes 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.

Inter-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.

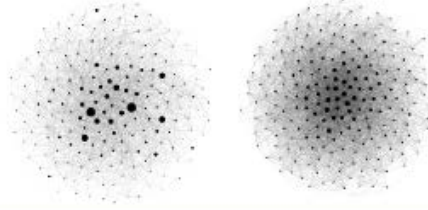
Internet neologisms

We obtained 80073 **Internet neologisms** Internet slangs from two online dictionary sources, HoSlang.com and Urban Dictionary.

Frequency	Neologisms
Frequent	lol, /s, kinda, bitcoin, idk, lmao, tbh, tl;dr, alot, /s, omg, lol, hahaha, irc, thugmonster, blein, soik, f'ang
Infrequent	yobbbish, fernanli, some, vampy

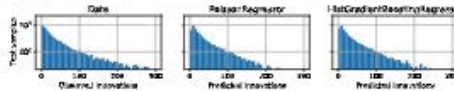
Selected Findings

After controlling for size, the network with **higher average degree** (more inner-connections) (right: $r/F13C, hegams$) 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.



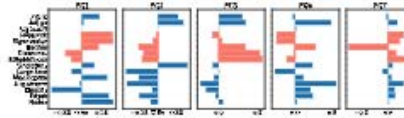
Lexical innovation

- 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.



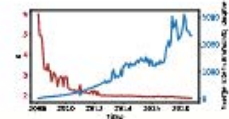
Lexical survival

- 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 connectness** also contribute to neologism survival.



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 or below**) despite increased inter-community connectedness (**increased average community degree below**)
- Segregation in topics and interests naturally brings in more community specific words.



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 levelling process in online communities with implications for sociolinguistic theories.

Acknowledgements

For work, Professor Patrick Ballew, Professor David Foray, John W. Burt Foster, Jr., David Foray, Adam Labaree and anonymous reviewers for their comments on earlier versions of this work. This material is based upon work supported by the National Science Foundation under Grant No. 1808222.

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**

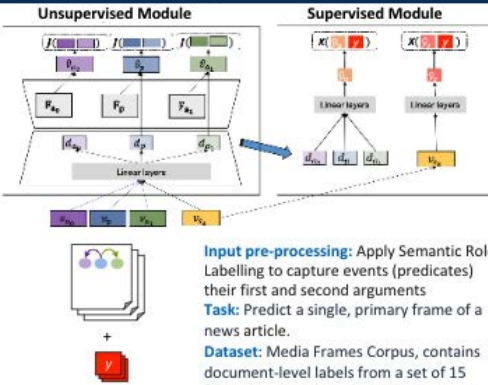
Intuition

- 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 **clarify the issue on the campaign trail** this fall. **The Arizona measure**, which was **upheld** in law by **Justice Roberts** 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** 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 that Democrats necessarily went along with the Arizona proposal** on the immigration issue. **They cite** the Obama administration's aggressive approach to fighting the Arizona law is yet more evidence of that out-of-touchness. In that vein, **nearly two dozen House Republicans sent letters** to Attorney General Eric Holder on Tuesday **describing the legal challenge** as the "height of irresponsibility and arrogance." **Polling on the Arizona law** specifically **tilts a Republican's favor**, although **broader data suggests a public** deeply **opposed** on immigration. In the latest Washington Post/ABC poll, **59 percent expressed support for the Arizona law** - including **42 percent who were strongly supportive** - while **33 percent opposed it**.

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

The Proposed Model (FRISS)

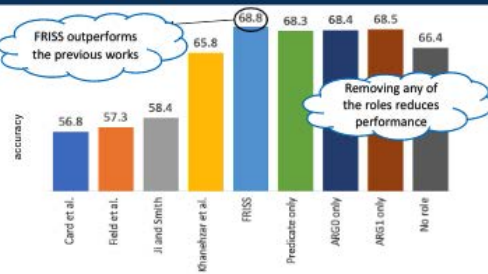


Input pre-processing: Apply Semantic Role Labelling to capture events (predicates) their first and second arguments

Task: Predict a single, primary frame of a news article.

Dataset: Media Frames Corpus, contains document-level labels from a set of 15 frames. covers several contagious issues, we focus on 'immigration'

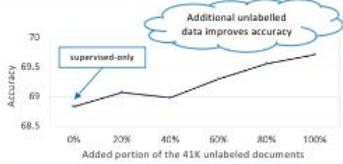
Experiment 1: Frame Prediction



FRISS outperforms the previous works

Removing any of the roles reduces performance

Experiment 2: Benefit of Unlabelled Data



Experiment 3: Qualitative Evaluation

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

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/shinyemialef/FRISS>

Modeling framing in immigration discourse on social media

1 Summary

We combine political communication and NLP to analyze the public's production and reception of frames in immigration discourse on Twitter

- We create a novel dataset of tweets labeled for multiple frame typologies
- We formulate frame detection as multilabel classification task
- Region and ideology influence framing
- Framing impacts audience responses

Joe Biden @JoeBiden · Apr 25
 Immigration was elected into our country against our diversity it, and has always been, but greatest in sept. Donald Trump doesn't get that—we need a president who does.

Democrats are the problem. They don't care about crime and civil legal immigrants, it's many "now" had they "now" to pour into and "Met" or "Country, like 10-15. They can't win on their narrative and it's so they view, it's like a story to solve.

3 Data Collection

- 2.6M English tweets, 2018-2019
 - With immigration term, e.g. immigrant, undocumented, illegals
- 4.5K tweets annotated for all frames explicitly cued
 - 80-10-10 train, dev, test split
- Infer users' regions (US, UK or EU) & ideology with existing tools (Larsson et al., 2019; Barberá 2015)

4 Frame Detection

- Multilabel classification layer for each frame type atop fine-tuned RoBERTa LM

Average F1 scores on cross-validated test set, separated by US authors' ideologies.

- Higher performance for conservatives suggests that they are more consistent than liberals in framing immigration

5 Analysis of Framing and its Effects

2 What is a Frame?

- "Selecting some aspects of a perceived reality and make them more salient in a communicating text" (Erssan, 1991)
- Issue-generic Policy**
 - Crime, morality, economic, political
- Issue-generic Narrative**
 - Episodic*: focus on specific actions, examples, case studies, or events
 - Thematic*: more generic views, placing story in broader social/political context
- Immigration-specific**
 - Victims (e.g. war, discrimination)
 - Heroes (e.g. economy, cultural diversity)
 - Threats (e.g. jobs, public safety)

All data, pretrained models and code available! <https://github.com/juliamendelsohn/framing>

Region: USA vs UK

Political Ideology

- USA:** Ideologically extreme frames (e.g. threat: public order, morality) most associated with USA
- UK:** more economy/labor, culture, global relationships
- Liberals:** immigrants as heroes and victims & prefer episodic frames
- Conservatives:** immigrants as threats & prefer thematic frames
- Issue-specific frames are most ideologically extreme & reveal differences otherwise obscured

Audience Responses

- Favorites:** cultural, human interest
- Retweets:** security, safety, political
- Clear narratives are important. Both *episodic* and *thematic* frames get higher engagement

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

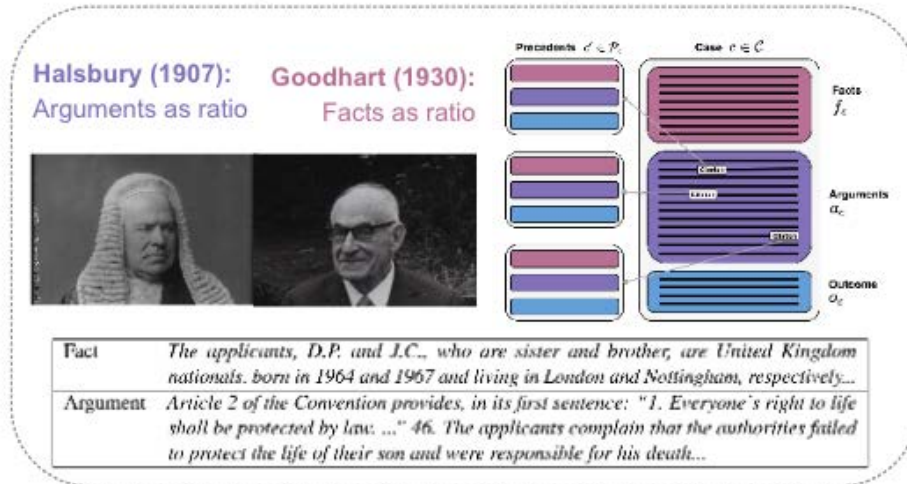
- Introduce the NT multilabel classification task for climate change scepticism
 - Labels: condemn (used to blame the alarmist greens), D-responsibility (used to highlight global warming being a natural cycle)

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.

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

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



$$MI(O; H | F) = H(O | F) - H(O | H, F)$$

$$MI(O; G | F) = H(O | F) - H(O | G, F)$$

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

Halsbury's Test

$p(o | h, f)$

Facts as ratio

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

Goodhart's Test

$p(o | g, f)$

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 t in subreddit s , its NPMI is:

$$NPMI(s, t) = \frac{\log \frac{P(s, t)}{P(s)P(t)}}{\log \frac{P(s, t)}{P(s)P(t)} + \log \frac{P(s, \bar{t})}{P(s)P(\bar{t})}}$$

Examples of words with high NPMI in a subreddit:



Word Senses

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

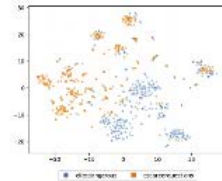
- Our word sense induction (WSI) method runs k-means directly on BERT embeddings.
- Amrami & Goldberg 2019's SOTA WSI method clusters word substitutes, which BERT predicts for a masked target word.

The SOTA model performs better on SemEval benchmarks, but our method is **4x 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..."

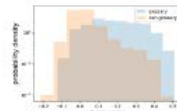
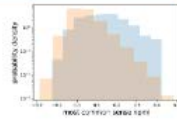


r/Elitedangerous: The [MASK] is a good multipurpose ship and a spectacular ship for grinding through missions. → blind, big, pilot, mormad, riper, pirates
r/Askarequests: No I've used [MASK], HTML, CSS, Javascript, node, flask.
→ slack, oracle, apple, bat, framework, windows
Top: t-SNE of BERT embeddings for pythons in two subreddits. Bottom: BERT substitutes for pythons.

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.

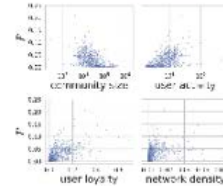


The distribution of scores for glossary words (blue) is higher than that for non-glossary words (orange)

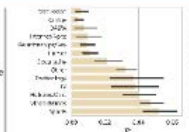
Community-level Attributes & Variation

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 loyal = 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 sociolect/style more often.
- Communities with **more dense** direct-reply networks



Left: The relationship between each community-level attribute and F_1 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_1 of communities in different topic categories. "DASW" stands for "Disgusting/Annoying/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 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
CU, LMU Munich, Germany · {schick,schuetze}@in.tum.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

_____ : Bat-winged dinosaurs were clumsy fliers.

$$p_i(\text{economist}) < p_i(\text{tics})$$

_____ : Bat-winged dinosaurs were clumsy fliers.

$$p_j(\text{economist})$$

$$p(1) = p_i(\text{tics}) - p_j(\text{economist})$$

TRAINING

_____ : Bat-winged dinosaurs were clumsy fliers.

$$P_{\text{MASK}1} \quad P_{\text{MASK}2}$$

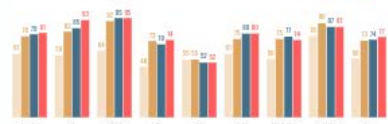
$$p(0) = p_{\text{MASK}1}(\text{science})$$

$$p(1) = p_{\text{MASK}1}(\text{economist}) - p_{\text{MASK}2}(\text{tics})$$

4 And How Well Does PET Work?

Results on SuperGLUE

with ALBERT-xxlarge-v2 for 32 examples

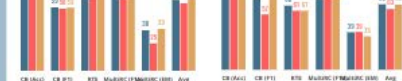


Different Patterns

using PET

WIC, WIC (QA), RTE

CB (QA), CB (PT), MultiRC (PT), MultiRC (QA), ANLI



Different Training Examples

using PET

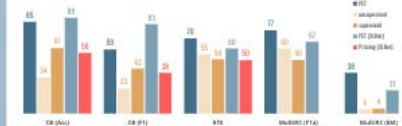
WIC, WIC (QA), RTE

CB (QA), CB (PT), MultiRC (PT), MultiRC (QA), ANLI



Different Few-Shot Methods

on selected SuperGLUE tasks



This work was funded by the European Research Council (ERC #101019636).

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 extent **factual knowledge**
- Obtain insights in the **underlying mechanism**
- Compare with **static embeddings**

Comparison

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

- ### Conclusions
- Static embeddings **competitive and cheap**
 - BERT great at **composing** representations from subwords
 - BERT considers **relation** information
 - Underlying mechanism** in BERT not more effective than NN-matching
 - 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
 e gets wordpiece tokenized into t_1, \dots, t_k
disregards relation information

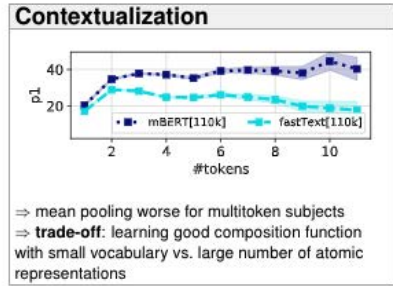
Advantages:

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

Multilingual Results

Model	Vocab. Size	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
fastText	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
	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



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 (bid) through their fellowship program. We thank Yanai Elazar and the anonymous reviewers for valuable comments.

- ### 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**

Presented at NAACL 2021, Online. * Equal Contribution

Session 11A (ethics)

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 $U(0,10000)$. For each seed pair, we compared the overall model performance (AUC) with performance for each subgroup (Δ AUC)
 - Δ AUC = $AUC_{subgroup} - AUC$

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)

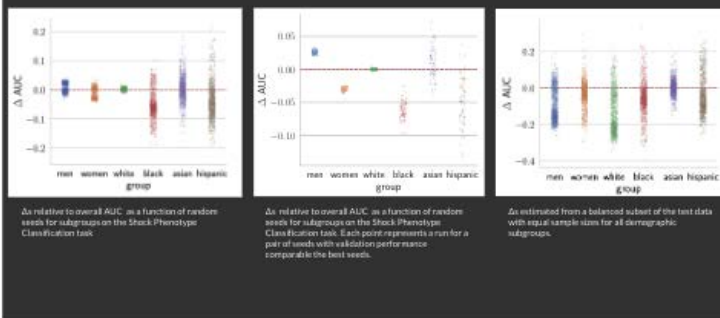


Silvio Amir, Jan-Willem van de Meent, Byron C. Wallace
 {s.amir, b.wallace, j.vandemeent}@northeastern.edu

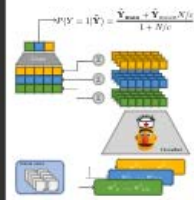
Random seeds significantly impact the performance and fairness of clinical classifiers on MIMIC-III

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

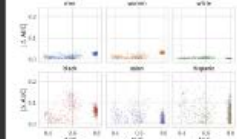
- model variability due to choice of **random seeds**
- variance due to **small sample sizes**



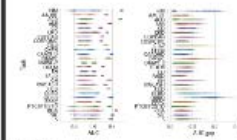
FINE-TUNED CLASSIFIERS



RESULTS



Correlations between overall performance and subgroup performance on the Shock Phenotype Classification task



Variation of model performance across random seeds. Left: Overall performance. Right: Gap between best and worst subgroup.

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



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 fits 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] Bessaki, A. et al. "Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering". In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[2] Stefania Drago et al. "Hey Google is it OK if I tell 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 Frames: Reassessing about Social and Power Implications of Language". In: *Proceedings of the 56th 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: I 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.

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.

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*₂.

The way forward: In order to avoid this, content moderation systems would need to make judgments based what is offensive rather than mere correlates. For example, [5] explicitly generate the implications of a statement which then could be used to judge offensiveness directly.

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]

- 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 to properly inform 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

- 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



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

Disparate model performance between different groups

Bias Factors

Group identifier bias / AAVE dialect bias / Gender bias

Advantages

- Significant reduction of efforts in downstream
- Encourages broader application of bias mitigation

Upstream Bias Mitigation (UBM) Framework

Setups

Task & Domain Similarity

- Same domain & task
- Cross domain & task

Number of Bias Factors Mitigated

- One bias factor
- Multiple bias factor

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

- 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

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

Privacy Regularization: Joint Privacy-Utility Optimization in Text-Generation Models

Fatemehsadat Mirehshgallah*, Huseyin Inan*, Marcello Hasegawa*, Victor Rühle*, Taylor Berg-Kirkpatrick*, Robert Sim*
¹UC San Diego ²Microsoft Research ³Microsoft

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

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. 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.

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

NAACL 2021 papers

39

Session 1E (sentence-level, textual inference)

Unifying cross-lingual SRL with heterogeneous linguistic resources

Unifying Cross-Lingual Semantic Role Labeling with Heterogeneous Linguistic Resources
Simone Conia, Andrea Bacchi and Roberto Navigli
Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome

NAACL 2021

Semantic Role Labeling (SRL)
A brief introduction
Semantic Role Labeling is the task of automatically identifying:

Who did What to Whom, Where, When, Why, and How?
(Charniak and Sag, 1996; Steedman et al., 2006)

The quick brown fox jumps over the lazy dog
MRB (Noun) jump (Verb) over the lazy dog (PP)

1. **Identify** all dependencies in the sentence
2. **Classify** the dependencies into the most granular semantic roles possible
3. **Aggregate** the dependencies into the most granular semantic roles possible
4. **Align** the dependencies across languages
5. **Aggregate** the dependencies into the most granular semantic roles possible

Model Overview
Learning from heterogeneous linguistic resources
Recent advances may be deeply rooted beyond the surface realizations that different languages have available:

- It is **tokens**: linguistic words (tokens, POS tags)
- It is **argument structures**: the syntactic structure (dependencies, syntactic trees)

Model Architecture: universal encoders
Universal sentence encoder and universal predicate-argument encoder
In the universal encoders, predictor and argument representations are shared across languages

Experiments on Dependency-based SRL (CoNLL-2009)

How does our approach fare?
Comparison with He et al., 2019
Our approach achieves state-of-the-art results across all tested languages in CoNLL-2009

Cross-Lingual SRL in low-data settings
Results when reducing the size of each dataset to 10%
The benefits of our cross-lingual approach are evident in low-data settings

1-Shot Cross-Lingual SRL
1-shot learning: 1 sentence / predicate-argument
Our approach performs strongly in 1-shot learning compared to state-of-the-art baselines

Analysis, Discussion and Takeaways

One model for many linguistic resources
Multiple arguments in a single forward pass
A single forward pass automatically compares different inventories and linguistic resources

Aligning meaning across inventories
What our model implicitly learns

Conclusion
Key takeaways:
- Cross-lingual SRL
- Aligning different resources
- SRL, especially in low-data

Meta-learning for domain generalization in semantic parsing

Cross-Domain Semantic Parsing

Train

database: concert singer

Show all countries and the number of singers in each country.

`SELECT Country, count(*) FROM Singer GROUP BY Country`

Test

database: farm

Please show the different statuses of cities and the average population of cities with each status.

`SELECT Status, avg(Population) FROM City GROUP BY Status`

Domain Generalization: a parser needs to generalize to unseen domains

Cross-Lingual Cross-Domain Semantic Parsing

Train

database: concert singer

每个国家有多少歌手

`SELECT Country, count(*) FROM Singer GROUP BY Country`

Test

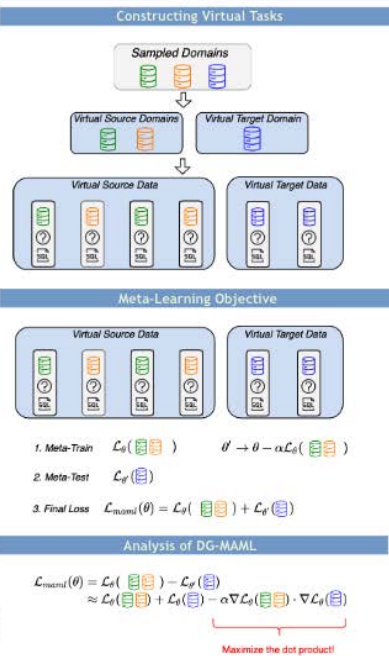
database: farm

请显示不同城市的状态和各个城市的城市平均人口。

`SELECT Status, avg(Population) FROM City GROUP BY Status`

- utterances and database schemas are in different languages
- ngram-based matching cannot be used for schema linking

DG-MAML

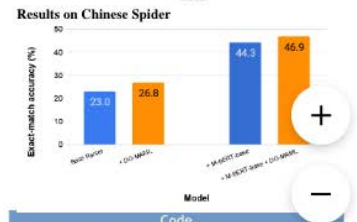
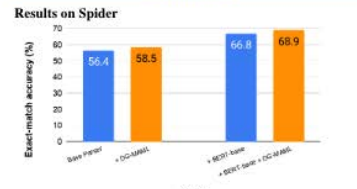


Gradient Updates of DG-MAML



DG-MAML encourages the gradients in virtual source and target domains to agree with each other

Experiments



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

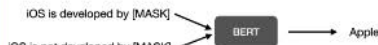
Session 4C (sentence-level, textual inference)

Understanding by understanding not: modeling negation in LMs

Abstract

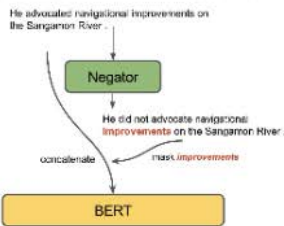
Negation is a core construction in natural language. Despite being very successful on many tasks, state-of-the-art pre-trained language models often handle negation incorrectly. To improve language models in this regard, we propose to augment the language modeling objective with an unlikelihood objective 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, NLI 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

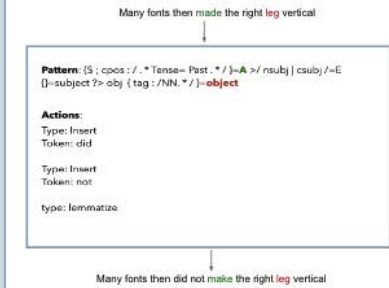
Unlikelihood with reference



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

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

Syntactic Negation Augmentation



Experimental Results

Model	SQuAD	ConceptNet	T-REx	Google-RE
BERT	13.53	15.65	25.10	10.24
BERT + XL	13.64	15.64	25.28	10.27
BERTNOT	13.97	15.49	25.25	10.31

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

Model	SQuAD	ConceptNet	T-REx	Google-RE
BERT	8.51	2.24	21.42	3.75
BERT + XL	4.97	1.19	21.77	3.93
BERTNOT	2.10	0.73	11.86	1.10

mean top 1 error rate for negated LAMA queries (lower is better)

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

Model	RTE		SNLI		+
	dev	w/mag	dev	w/mag	
BERT	78.4 _{±0.8}	65.47 _{±0.15}	89.47 _{±0.19}	64.18 _{±0.07}	64.10
BERTNOT	69.8 _{±1.0}	74.47 _{±0.39}	89.0 _{±0.19}	45.9 _{±0.07}	60.89 _{±0.08}

Annotations on original dev splits and new splits containing negation from Hosseini et al. (2020) w/mag

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

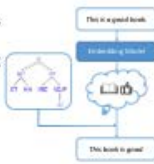
Motivation

- **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

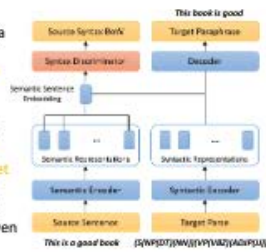


ParaBART: Learning Disentanglement from Paraphrases

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

$$\min_{\theta} \mathbb{E}_{s \sim P(s)} \mathbb{E}_{t \sim P(t|s)} \left(\max_{\phi} (\mathcal{L}_{\text{para}} - \lambda_{\text{dis}} \mathcal{L}_{\text{dis}}) \right)$$

Paraphrasing Model Syntax Discriminator

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	STS22	STS19	STS28	STS16	STS18	STS-8	Avg.
Avg. BERT embeddings	86.0	82.8	87.2	83.5	84.5	87.8	85.5
Avg. BART embeddings	89.8	82.8	84.1	83.9	88.5	82.6	84.7
inferSent	89.2	88.8	89.0	71.5	71.5	78.8	80.9
MGMC	41.8	62.3	68.2	72.5	67.8	78.2	68.0
Universal Sentence Encoder	81.8	88.5	89.6	76.3	71.9	78.2	80.7
Sentence-BERT	84.6	87.5	88.2	78.3	70.1	78.1	80.6
ParaBART	88.9	82.2	85.9	79.5	79.3	—	—
ParaBART	88.9	77.2	78.1	80.1	80.1	81.9	79.4
- with Adversarial Loss	87.5	79.8	75.8	80.9	80.0	78.3	75.5
- with Adversarial Loss & Syntactic Guidance	86.4	81.9	73.8	80.0	79.8	78.4	78.2

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



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

1. Contribution

TRACIE (Temporal 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 Labeling

- 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.

Head

I went to the park on January 1st. I was very hungry after some hiking. Luckily, I purchased a lot of food before I went to the park. I enjoyed the trip and wrote an online review about the trip in the 10th.

within sentence before

I purchased food. I went to the park.

cross-sentence before

I went to the park. I wrote a review.

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

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 $sign(duration_2 - (start_2 - start_1))$

Sub-modules.

- PatternTime provides $start_2 - start_1$.
- For $duration_2$, 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 ϕ to get a weighted mean $f(x)$. We use the binary label from PatternTime and apply a $tanh$ function to get a sign close to either -1 or 1 from probabilities $\{g(x)\}$. 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**

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 T5-3B, showing the benefit of distant supervision + reasoning

SymTime-ZS outperforms existing pre-trained and finetuned language models.

- Smaller models can be more efficient than large ones, with distant supervision and correct reasoning processes.

We also see that all baselines' performance drop significantly when the distribution priors are removed from the training data. However, our models are more consistent across settings.

(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?

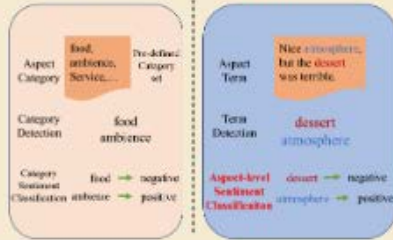
A strong baseline for Aspect-based Sentiment Analysis with RoBERTa



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

INTRODUCTION

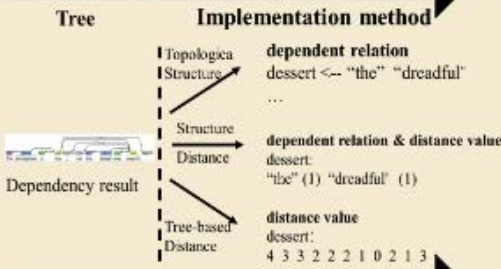
The atmosphere is nice, but the dessert was dreadful.



Aspect Category		Aspect Term	
food	ambience	dessert	atmosphere
negative	positive	negative	positive

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



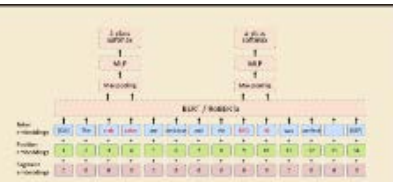
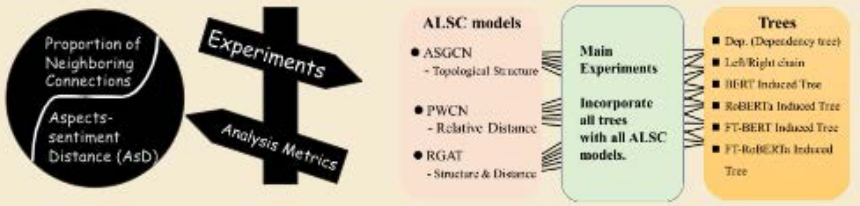
Questions:

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



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

PTMs and TREES



Conclusions

• A Simple MLP based on RoBERTa COULD OBTAIN THE SOTA IN ALSC.
RoBERTa, YES!

Acknowledgment

National Natural Science Foundation of China [No. 62022027]
National Key Research and Development Program of China [No. 2020AAA0106700]

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}

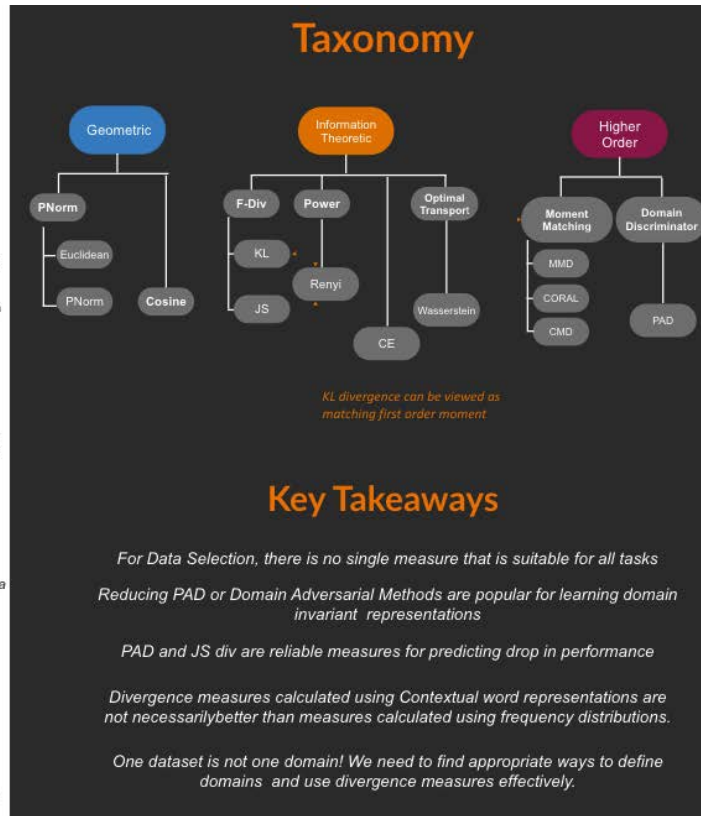
Correlation of performance drop with the divergence between domains

Datasets

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

Divergence Measures

- 12 divergence measures used in the literature

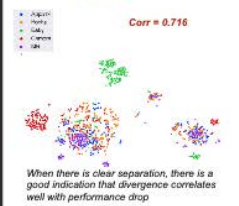


Divergence/Task	POS	NER	SA
Cos	0.018	0.223	-0.012
KL-Div	0.394	0.384	0.716
JS-Div	0.407	0.484	0.709
Renyi Div	0.392	0.382	0.716
PAD	0.477	0.426	0.538
Wasserstein	0.378	0.463	0.448
MMD-RQ	0.248	0.495	0.614
MMD-Gaussian	0.402	0.221	0.543
MMD-Energy	0.244	0.447	0.521
MMD-Laplacian	0.389	0.273	0.623
CORAL	0.340	0.484	0.267

Correlation between divergence Measure and performance drop

Divergence/Tasks	POS	NER	SA
Cos	-1.78 x 10 ⁻¹	-2.43 x 10 ⁻¹	-3.01 x 10 ⁻¹
KL-Div	-	-	-
JS-Div	-4.3 x 10 ⁻¹	-4.4 x 10 ⁻¹	3.84 x 10 ⁻¹
Renyi Div	-	-	-
PAD	-	-	-
Wasserstein	-2.11 x 10 ⁻¹	-2.36 x 10 ⁻¹	-1.70 x 10 ⁻¹
MMD-RQ	-4.11 x 10 ⁻¹	-3.84 x 10 ⁻¹	-1.70 x 10 ⁻¹
MMD-Gaussian	4.26 x 10 ⁻¹	2.37 x 10 ⁻¹	-8.45 x 10 ⁻¹
MMD-Energy	-8.46 x 10 ⁻¹	-1.14 x 10 ⁻¹	-2.48 x 10 ⁻¹
MMD-Laplacian	-1.67 x 10 ⁻¹	4.25 x 10 ⁻¹	-1.08 x 10 ⁻¹
CORAL	-2.34 x 10 ⁻¹	-2.18 x 10 ⁻¹	-1.41 x 10 ⁻¹

Silhouette coefficients with different divergence measures.
Dataset-is-not-a-domain



Session 8C (sentence-level, textual inference)

Learning from executions for semantic parsing

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	Exec
SELECT restaurant WHERE star rating = 3	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

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

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_{UN}
Non-Executable Programs	P_{SN}	P_{IN}

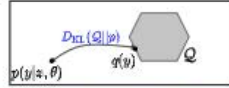
Beam search can help us see a subset of programs

Posterior Regularization

We assume a constrained family of distribution Q : for any $q \in Q$,

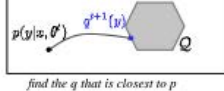
$$\mathbb{E}_{q(y)}[R(y)] = 1$$

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

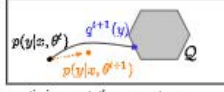


EM Algorithm for Optimizing PR

E-step



M-step



Optimizing PR is equivalent to optimizing MML!

New Interpretation of Conventional Methods

Self-Training:

$$q_{ST}^{t+1}(y) = \begin{cases} 1 & y = y^* \\ 0 & \text{otherwise} \end{cases}$$

Top-K MML:

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

New Objectives

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

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

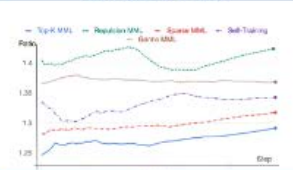
$$q_{sparse}^{t+1} = \text{SparseMax}_{y \in P_{UN}} (\log p(y|x, \theta^t))$$

Experiments

Results on Overnight



Analysis: Length Ratio



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

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

Semantic Parsing: Natural Language Interfaces for Computers

Schedule a meeting with Jean at 5.

Who is Jean's manager?

Who is on Abby's team?

Compositional Generalization for Semantic Parsing

Research Question: How to generalize a neural semantic parser to understand compositionally novel utterances not encountered during training time?

Contributions: A simple supervised attention method that works with LSTMs and Transformers. A new realistic benchmark dataset for evaluating compositional generalization.

Training Examples: Utterances of Single-domain Skills

Testing Examples: Compositional Utterances Across Multiple Domains

Span-level Supervised Attention between Utterance and Program Spans

Alignments between Utterance and Program Spans

Intuition: Encourage a neural attentional program decoder to make predictions of sub-programs using localized information of the aligned utterance spans

Generate word-level alignments using IBM Model 4

Leverage program structures to expand word-level alignments to span-level alignments

Span-level Supervised Attention in Attentional Neural Decoders

Regularize the neural decoder's target-to-source attention distribution according to the pre-defined alignments between sub-programs $y_{i,k}$ and utterance spans $u_{i,j}$

$$P_{att}(u_{i,j} = y_{i,k}) = \frac{1}{|u_{i,j}|} \text{ if } u \in u_{i,j} \text{ else } 0$$

CaLiFlow Compositional Skills Benchmark Dataset

Training Single-Skill Utterances

Testing Compositional-Skill Utterances

Real-world examples: manually curated utterances for calendar management and orchard query, extracted from the Microsoft CaLiFlow task-oriented dialogue set (Semantic Machines et al., 2020).

Few-shot learning setting: a handful of compositional-skill examples are used for training.

EXPERIMENTS

Accuracies on Predicting Compositionally Novel Examples

Visualization of Attention Distribution

Observation: Span-level supervised attention yields more structured attention distribution

Results on Compositional Freebase Questions (Keyzers et al., 2020)

Conjunctive Examples

Recursive Examples (multi-hop questions)

Span-level attention supervision is more effective on recursive questions

Incorporating external knowledge to enhance tabular reasoning

Incorporating External Knowledge to Enhance Tabular Reasoning

J. Neeraja⁽¹⁾, Vivek Gupta⁽²⁾, Vivek Srikumar⁽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
- 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
- Exposes 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 cropped.

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

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.

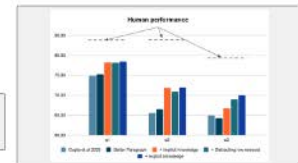
Approach

- 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:      $\pi \leftarrow \text{Random-Permutation}(V)$ 
6:      $C_1 \leftarrow b(w, \pi)$  (predecessors of  $w$ )
7:      $C_2 \leftarrow C_1 \cup \{w\}$  (predecessors including  $w$ )
8:      $f_w^{C_1} \leftarrow \text{TrainModel}(C_1)$  (Train on vocabulary  $C_1$ )
9:      $f_w^{C_2} \leftarrow \text{TrainModel}(C_2)$  (Train on vocabulary  $C_2$ )
10:     $m(w, \pi) \leftarrow q(f_w^{C_2}) - q(f_w^{C_1})$ 
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
    
```

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

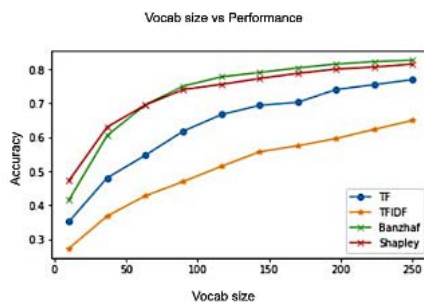


Figure 1. Comparing Shapley, Banzhaf, TF (term frequency) and TF-IDF on AG-news (a document classification task). We see that both game-theoretic algorithms (red and green) outperform the other heuristics for all vocabulary sizes.

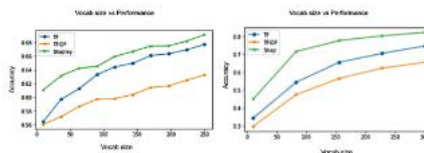


Figure 2. Comparison of Shapley to TF (frequency) and TF-IDF on two additional task structures: single-sentence classification (left) on SST-2 and pairwise-sentence classification (right) on SNLI. We see that Shapley (green) outperforms the baselines in both tasks here, as well as the third task structure in Fig 1.

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

FLIN: A Flexible Natural Language Interface for Web Navigation

Sahisnu Mazumder¹ and Oriana Riva²

¹Department of Computer Science, University of Illinois at Chicago, USA

²Microsoft Research, Redmond, USA

sahisnumazumder@gmail.com, oriana.riva@microsoft.com

NAACL 2021



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 maps 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 semantic parsing methods deal with environments that have a fixed and known set of actions, which is not the case with real websites.
- The same concept-level action can have different logical representations and parameter schema 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

Two sub-tasks:

- Action recognition
- Parameter recognition and value assignment

Dealing with closed-domain parameters

Extract parameter mention from the command and map it to correct parameter value.

Dealing with open-domain parameters

Extracted parameter mention is the value to be assigned. No mapping is needed.

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

Evaluation Setup

- WebNav dataset:** A labelled dataset consisting of 9 websites from Restaurants (R), Hotels (H) and Shopping (S) domains.
- DialQueries dataset:** A real user query dataset (R: 421, H: 155, S: 63) adapted from SGD dialogue and Dialogflow.

# websites	# actions	# parameters	Dataset splits
Restaurants (R)	4	26	14332 / 2865 / 1913
Hotels (H)	7	34	19
Shopping (S)	7	34	19
Dialogflow	7	34	19
SGD	7	34	19
Microsoft Research	7	34	19
Microsoft Research	7	34	19
Microsoft Research	7	34	19
Microsoft Research	7	34	19
Microsoft Research	7	34	19

A Flexible Web Navigation System

Once the Action Extractor and Action Executor modules are built, they can be maintained automatically. Building a flexible semantic parser can make the whole system scale.

FLIN Architecture

Experimental Results

FLIN and its variants adapt well to previously-unseen websites

Website	Acc	F1	EMA	PA-100
opendata.com (R)	0.925	0.899	0.906	0.910
reel.com (R)	0.810	0.790	0.823	0.808
reel.com (S)	0.839	0.824	0.843	0.869

In-website & cross-website performance comparison on WebNav dataset

Website	Acc	F1	EMA	PA-100
opendata.com (R)	0.719	0.582	0.565	0.562
hotels.com (H)	0.730	0.514	0.281	0.560
reel.com (S)	0.507	0.464	0.428	0.440

Performance of FLIN on DialQueries dataset

Command	Acc	F1	EMA	PA-100
Gold: filter by category (category: kids footwear)	17.7	8.9	13.5	13.5
Gold: filter by gender (gender: kids)	17.7	8.9	13.5	13.5
Gold: find table (party size: 3)	21.0	10.7	14.1	14.1
Gold: search (search within the reviews: "me my wife and my daughter")	31.1	6.1	14.1	14.1
Gold: booking time (booking time: 20:00)	15.5	11.3	16.4	17.0

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

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

- 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)

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 is more or less at the same level as MTL's

Results: Rules and MTL on each rule category

- Rules outperforms MTL on majority of categories
- MTL achieves SOTA by resolving anaphors in the largest category, Rule 18 (anaphors that cannot be handled by any of the rules)
- Rules outperforms MTL on less categories than in ISNotes
- BASHI has lower resolution precision than Rules in ISNotes even though the rulesets used for ISNotes and BASHI are almost identical
- Rules outperforms MTL on only two of the rule categories

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

Is incoherence surprising? Targeted evaluation of coherence prediction from LMs

Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models



Anne Beyer, Sharid Loáiciga, David Schlangen
 Computational Linguistics, Department of Linguistics, University of Potsdam
 anne.beyer@uni-potsdam.de



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

Previous Work

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

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						
	intro	rp subj	prep	the	prep rp	matrix verb	continuation
match_sing	The	farmer	saw	the	clerk	knows	many people
mismatch_sing	The	farmer	saw	the	clerk	knows	many people
mismatch_plural	The	farmers	saw	the	clerk	knows	many people
mismatch_gram	The	farmers	saw	the	clerk	knows	many people

Item 1

match_sing	The	manager	to the side of	the	architect	likes	to gamble
mismatch_sing	The	manager	to the side of	the	architect	like	to gamble
mismatch_plural	The	managers	to the side of	the	architect	likes	to gamble
mismatch_gram	The	managers	to the side of	the	architect	like	to gamble

Item 2

Prediction: $\{ \text{match_sing} \text{ matrix verb } < \text{mismatch_sing} \text{ matrix verb} \}$
 & $\{ \text{match_plural} \text{ matrix verb } < \text{mismatch_plural} \text{ matrix verb} \}$

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**

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:

CONNECTIVE SENSE	GPT-2						
	although	as	however	since	though	while	yet
as_causal	0.44	–	0.80	0.28	0.64	0.72	0.80
as_comparative	0.96	–	0.95	0.96	0.94	0.97	0.97
as_concessive	0.33	–	0.57	0.57	0.33	1.00	1.00
as_preposition	0.00	–	0.94	0.96	0.95	0.93	0.96
as_temporal	0.95	–	0.95	0.96	1.00	0.81	0.95

Speaker Commitment

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

context: A: since the beginning of the year, I am a nurse.
 original: **B: 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 LMs

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

- Investigated inner working of pre-trained transformer based language models, specifically for **Bridging** information.
- Presented two approaches of investigation:
 - Probing of individual attention heads
 - Of-cloze task to examine whole model

FINDINGS

- 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.

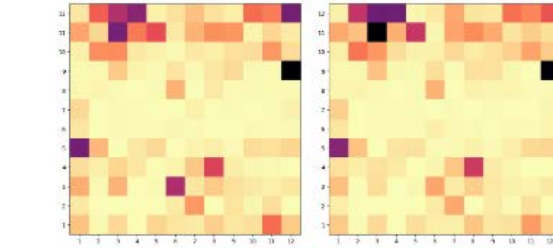


Figure 1: Bridging signals with BERT-base-based model where anaphor and antecedents are 2 sentences apart. Bridging signals from anaphor to antecedent are shown in the first heatmap and the reverse signals in the second. In both heatmaps, the x-axis shows the attention head number and the y-axis shows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

1. *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
- Comparison with result obtained with prominent heads.
2. *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 "of"	without "of"	perturb
only anaphor	17.20	5.62	-
ana sent.	22.82	7.71	10.28
ana+ante sent.	27.81	9.61	10.93
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-based 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*.


Universal discourse representation structure parsing



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) is 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.
- DRT 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.

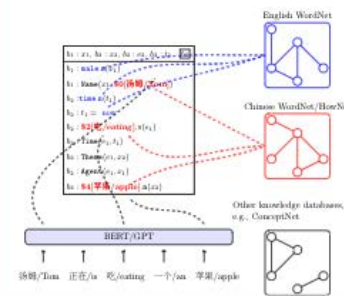
pre-supposition pointer scope-binding operator link labels

$A_1: x_1, y_1, z_1, w_1$ $A_2: x_2, y_2, z_2, w_2$ $A_3: x_3, y_3, z_3, w_3$ $A_4: x_4, y_4, z_4, w_4$ $A_5: x_5, y_5, z_5, w_5$ $A_6: x_6, y_6, z_6, w_6$ $A_7: x_7, y_7, z_7, w_7$ $A_8: x_8, y_8, z_8, w_8$ $A_9: x_9, y_9, z_9, w_9$ $A_{10}: x_{10}, y_{10}, z_{10}, w_{10}$ $A_{11}: x_{11}, y_{11}, z_{11}, w_{11}$ $A_{12}: x_{12}, y_{12}, z_{12}, w_{12}$ $A_{13}: x_{13}, y_{13}, z_{13}, w_{13}$ $A_{14}: x_{14}, y_{14}, z_{14}, w_{14}$ $A_{15}: x_{15}, y_{15}, z_{15}, w_{15}$ $A_{16}: x_{16}, y_{16}, z_{16}, w_{16}$ $A_{17}: x_{17}, y_{17}, z_{17}, w_{17}$ $A_{18}: x_{18}, y_{18}, z_{18}, w_{18}$ $A_{19}: x_{19}, y_{19}, z_{19}, w_{19}$ $A_{20}: x_{20}, y_{20}, z_{20}, w_{20}$ $A_{21}: x_{21}, y_{21}, z_{21}, w_{21}$ $A_{22}: x_{22}, y_{22}, z_{22}, w_{22}$ $A_{23}: x_{23}, y_{23}, z_{23}, w_{23}$ $A_{24}: x_{24}, y_{24}, z_{24}, w_{24}$ $A_{25}: x_{25}, y_{25}, z_{25}, w_{25}$ $A_{26}: x_{26}, y_{26}, z_{26}, w_{26}$ $A_{27}: x_{27}, y_{27}, z_{27}, w_{27}$ $A_{28}: x_{28}, y_{28}, z_{28}, w_{28}$ $A_{29}: x_{29}, y_{29}, z_{29}, w_{29}$ $A_{30}: x_{30}, y_{30}, z_{30}, w_{30}$ $A_{31}: x_{31}, y_{31}, z_{31}, w_{31}$ $A_{32}: x_{32}, y_{32}, z_{32}, w_{32}$ $A_{33}: x_{33}, y_{33}, z_{33}, w_{33}$ $A_{34}: x_{34}, y_{34}, z_{34}, w_{34}$ $A_{35}: x_{35}, y_{35}, z_{35}, w_{35}$ $A_{36}: x_{36}, y_{36}, z_{36}, w_{36}$ $A_{37}: x_{37}, y_{37}, z_{37}, w_{37}$ $A_{38}: x_{38}, y_{38}, z_{38}, w_{38}$ $A_{39}: x_{39}, y_{39}, z_{39}, w_{39}$ $A_{40}: x_{40}, y_{40}, z_{40}, w_{40}$ $A_{41}: x_{41}, y_{41}, z_{41}, w_{41}$ $A_{42}: x_{42}, y_{42}, z_{42}, w_{42}$ $A_{43}: x_{43}, y_{43}, z_{43}, w_{43}$ $A_{44}: x_{44}, y_{44}, z_{44}, w_{44}$ $A_{45}: x_{45}, y_{45}, z_{45}, w_{45}$ $A_{46}: x_{46}, y_{46}, z_{46}, w_{46}$ $A_{47}: x_{47}, y_{47}, z_{47}, w_{47}$ $A_{48}: x_{48}, y_{48}, z_{48}, w_{48}$ $A_{49}: x_{49}, y_{49}, z_{49}, w_{49}$ $A_{50}: x_{50}, y_{50}, z_{50}, w_{50}$ $A_{51}: x_{51}, y_{51}, z_{51}, w_{51}$ $A_{52}: x_{52}, y_{52}, z_{52}, w_{52}$ $A_{53}: x_{53}, y_{53}, z_{53}, w_{53}$ $A_{54}: x_{54}, y_{54}, z_{54}, w_{54}$ $A_{55}: x_{55}, y_{55}, z_{55}, w_{55}$ $A_{56}: x_{56}, y_{56}, z_{56}, w_{56}$ $A_{57}: x_{57}, y_{57}, z_{57}, w_{57}$ $A_{58}: x_{58}, y_{58}, z_{58}, w_{58}$ $A_{59}: x_{59}, y_{59}, z_{59}, w_{59}$ $A_{60}: x_{60}, y_{60}, z_{60}, w_{60}$ $A_{61}: x_{61}, y_{61}, z_{61}, w_{61}$ $A_{62}: x_{62}, y_{62}, z_{62}, w_{62}$ $A_{63}: x_{63}, y_{63}, z_{63}, w_{63}$ $A_{64}: x_{64}, y_{64}, z_{64}, w_{64}$ $A_{65}: x_{65}, y_{65}, z_{65}, w_{65}$ $A_{66}: x_{66}, y_{66}, z_{66}, w_{66}$ $A_{67}: x_{67}, y_{67}, z_{67}, w_{67}$ $A_{68}: x_{68}, y_{68}, z_{68}, w_{68}$ $A_{69}: x_{69}, y_{69}, z_{69}, w_{69}$ $A_{70}: x_{70}, y_{70}, z_{70}, w_{70}$ $A_{71}: x_{71}, y_{71}, z_{71}, w_{71}$ $A_{72}: x_{72}, y_{72}, z_{72}, w_{72}$ $A_{73}: x_{73}, y_{73}, z_{73}, w_{73}$ $A_{74}: x_{74}, y_{74}, z_{74}, w_{74}$ $A_{75}: x_{75}, y_{75}, z_{75}, w_{75}$ $A_{76}: x_{76}, y_{76}, z_{76}, w_{76}$ $A_{77}: x_{77}, y_{77}, z_{77}, w_{77}$ $A_{78}: x_{78}, y_{78}, z_{78}, w_{78}$ $A_{79}: x_{79}, y_{79}, z_{79}, w_{79}$ $A_{80}: x_{80}, y_{80}, z_{80}, w_{80}$ $A_{81}: x_{81}, y_{81}, z_{81}, w_{81}$ $A_{82}: x_{82}, y_{82}, z_{82}, w_{82}$ $A_{83}: x_{83}, y_{83}, z_{83}, w_{83}$ $A_{84}: x_{84}, y_{84}, z_{84}, w_{84}$ $A_{85}: x_{85}, y_{85}, z_{85}, w_{85}$ $A_{86}: x_{86}, y_{86}, z_{86}, w_{86}$ $A_{87}: x_{87}, y_{87}, z_{87}, w_{87}$ $A_{88}: x_{88}, y_{88}, z_{88}, w_{88}$ $A_{89}: x_{89}, y_{89}, z_{89}, w_{89}$ $A_{90}: x_{90}, y_{90}, z_{90}, w_{90}$ $A_{91}: x_{91}, y_{91}, z_{91}, w_{91}$ $A_{92}: x_{92}, y_{92}, z_{92}, w_{92}$ $A_{93}: x_{93}, y_{93}, z_{93}, w_{93}$ $A_{94}: x_{94}, y_{94}, z_{94}, w_{94}$ $A_{95}: x_{95}, y_{95}, z_{95}, w_{95}$ $A_{96}: x_{96}, y_{96}, z_{96}, w_{96}$ $A_{97}: x_{97}, y_{97}, z_{97}, w_{97}$ $A_{98}: x_{98}, y_{98}, z_{98}, w_{98}$ $A_{99}: x_{99}, y_{99}, z_{99}, w_{99}$ $A_{100}: x_{100}, y_{100}, z_{100}, w_{100}$	$A_1: x_1, y_1, z_1, w_1$ $A_2: x_2, y_2, z_2, w_2$ $A_3: x_3, y_3, z_3, w_3$ $A_4: x_4, y_4, z_4, w_4$ $A_5: x_5, y_5, z_5, w_5$ $A_6: x_6, y_6, z_6, w_6$ $A_7: x_7, y_7, z_7, w_7$ $A_8: x_8, y_8, z_8, w_8$ $A_9: x_9, y_9, z_9, w_9$ $A_{10}: x_{10}, y_{10}, z_{10}, w_{10}$ $A_{11}: x_{11}, y_{11}, z_{11}, w_{11}$ $A_{12}: x_{12}, y_{12}, z_{12}, w_{12}$ $A_{13}: x_{13}, y_{13}, z_{13}, w_{13}$ $A_{14}: x_{14}, y_{14}, z_{14}, w_{14}$ $A_{15}: x_{15}, y_{15}, z_{15}, w_{15}$ $A_{16}: x_{16}, y_{16}, z_{16}, w_{16}$ $A_{17}: x_{17}, y_{17}, z_{17}, w_{17}$ $A_{18}: x_{18}, y_{18}, z_{18}, w_{18}$ $A_{19}: x_{19}, y_{19}, z_{19}, w_{19}$ $A_{20}: x_{20}, y_{20}, z_{20}, w_{20}$ $A_{21}: x_{21}, y_{21}, z_{21}, w_{21}$ $A_{22}: x_{22}, y_{22}, z_{22}, w_{22}$ $A_{23}: x_{23}, y_{23}, z_{23}, w_{23}$ $A_{24}: x_{24}, y_{24}, z_{24}, w_{24}$ $A_{25}: x_{25}, y_{25}, z_{25}, w_{25}$ $A_{26}: x_{26}, y_{26}, z_{26}, w_{26}$ $A_{27}: x_{27}, y_{27}, z_{27}, w_{27}$ $A_{28}: x_{28}, y_{28}, z_{28}, w_{28}$ $A_{29}: x_{29}, y_{29}, z_{29}, w_{29}$ $A_{30}: x_{30}, y_{30}, z_{30}, w_{30}$ $A_{31}: x_{31}, y_{31}, z_{31}, w_{31}$ $A_{32}: x_{32}, y_{32}, z_{32}, w_{32}$ $A_{33}: x_{33}, y_{33}, z_{33}, w_{33}$ $A_{34}: x_{34}, y_{34}, z_{34}, w_{34}$ $A_{35}: x_{35}, y_{35}, z_{35}, w_{35}$ $A_{36}: x_{36}, y_{36}, z_{36}, w_{36}$ $A_{37}: x_{37}, y_{37}, z_{37}, w_{37}$ $A_{38}: x_{38}, y_{38}, z_{38}, w_{38}$ $A_{39}: x_{39}, y_{39}, z_{39}, w_{39}$ $A_{40}: x_{40}, y_{40}, z_{40}, w_{40}$ $A_{41}: x_{41}, y_{41}, z_{41}, w_{41}$ $A_{42}: x_{42}, y_{42}, z_{42}, w_{42}$ $A_{43}: x_{43}, y_{43}, z_{43}, w_{43}$ $A_{44}: x_{44}, y_{44}, z_{44}, w_{44}$ $A_{45}: x_{45}, y_{45}, z_{45}, w_{45}$ $A_{46}: x_{46}, y_{46}, z_{46}, w_{46}$ $A_{47}: x_{47}, y_{47}, z_{47}, w_{47}$ $A_{48}: x_{48}, y_{48}, z_{48}, w_{48}$ $A_{49}: x_{49}, y_{49}, z_{49}, w_{49}$ $A_{50}: x_{50}, y_{50}, z_{50}, w_{50}$ $A_{51}: x_{51}, y_{51}, z_{51}, w_{51}$ $A_{52}: x_{52}, y_{52}, z_{52}, w_{52}$ $A_{53}: x_{53}, y_{53}, z_{53}, w_{53}$ $A_{54}: x_{54}, y_{54}, z_{54}, w_{54}$ $A_{55}: x_{55}, y_{55}, z_{55}, w_{55}$ $A_{56}: x_{56}, y_{56}, z_{56}, w_{56}$ $A_{57}: x_{57}, y_{57}, z_{57}, w_{57}$ $A_{58}: x_{58}, y_{58}, z_{58}, w_{58}$ $A_{59}: x_{59}, y_{59}, z_{59}, w_{59}$ $A_{60}: x_{60}, y_{60}, z_{60}, w_{60}$ $A_{61}: x_{61}, y_{61}, z_{61}, w_{61}$ $A_{62}: x_{62}, y_{62}, z_{62}, w_{62}$ $A_{63}: x_{63}, y_{63}, z_{63}, w_{63}$ $A_{64}: x_{64}, y_{64}, z_{64}, w_{64}$ $A_{65}: x_{65}, y_{65}, z_{65}, w_{65}$ $A_{66}: x_{66}, y_{66}, z_{66}, w_{66}$ $A_{67}: x_{67}, y_{67}, z_{67}, w_{67}$ $A_{68}: x_{68}, y_{68}, z_{68}, w_{68}$ $A_{69}: x_{69}, y_{69}, z_{69}, w_{69}$ $A_{70}: x_{70}, y_{70}, z_{70}, w_{70}$ $A_{71}: x_{71}, y_{71}, z_{71}, w_{71}$ $A_{72}: x_{72}, y_{72}, z_{72}, w_{72}$ $A_{73}: x_{73}, y_{73}, z_{73}, w_{73}$ $A_{74}: x_{74}, y_{74}, z_{74}, w_{74}$ $A_{75}: x_{75}, y_{75}, z_{75}, w_{75}$ $A_{76}: x_{76}, y_{76}, z_{76}, w_{76}$ $A_{77}: x_{77}, y_{77}, z_{77}, w_{77}$ $A_{78}: x_{78}, y_{78}, z_{78}, w_{78}$ $A_{79}: x_{79}, y_{79}, z_{79}, w_{79}$ $A_{80}: x_{80}, y_{80}, z_{80}, w_{80}$ $A_{81}: x_{81}, y_{81}, z_{81}, w_{81}$ $A_{82}: x_{82}, y_{82}, z_{82}, w_{82}$ $A_{83}: x_{83}, y_{83}, z_{83}, w_{83}$ $A_{84}: x_{84}, y_{84}, z_{84}, w_{84}$ $A_{85}: x_{85}, y_{85}, z_{85}, w_{85}$ $A_{86}: x_{86}, y_{86}, z_{86}, w_{86}$ $A_{87}: x_{87}, y_{87}, z_{87}, w_{87}$ $A_{88}: x_{88}, y_{88}, z_{88}, w_{88}$ $A_{89}: x_{89}, y_{89}, z_{89}, w_{89}$ $A_{90}: x_{90}, y_{90}, z_{90}, w_{90}$ $A_{91}: x_{91}, y_{91}, z_{91}, w_{91}$ $A_{92}: x_{92}, y_{92}, z_{92}, w_{92}$ $A_{93}: x_{93}, y_{93}, z_{93}, w_{93}$ $A_{94}: x_{94}, y_{94}, z_{94}, w_{94}$ $A_{95}: x_{95}, y_{95}, z_{95}, w_{95}$ $A_{96}: x_{96}, y_{96}, z_{96}, w_{96}$ $A_{97}: x_{97}, y_{97}, z_{97}, w_{97}$ $A_{98}: x_{98}, y_{98}, z_{98}, w_{98}$ $A_{99}: x_{99}, y_{99}, z_{99}, w_{99}$ $A_{100}: x_{100}, y_{100}, z_{100}, w_{100}$
--	--

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.

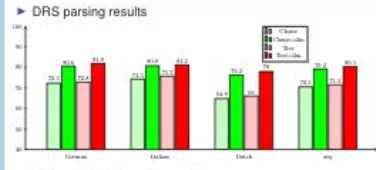


BERT/OPPT

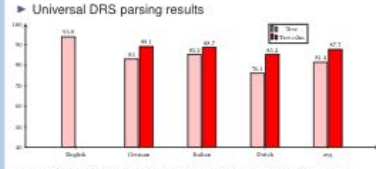
汤姆/Tom 正在/is 吃/eating 一个/a 苹果/apple

Low-resource language Experiments on the Parallel Meaning Bank

► DRS parsing results



► Universal DRS parsing results



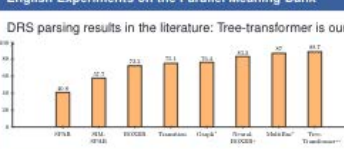
► We construct Universal DRS dataset for **more than 100** languages

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

English Experiments on the Parallel Meaning Bank

DRS parsing results in the literature: Tree-transformer is ours



The constructed Universal DRS quality

Auto-generated vs gold Universal DRS

language	BLEU1	F1
de	45.03	94.25
it	62.22	89.43
nl	69.12	94.06
avg	65.12 (±3.0)	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	nl	avg
correct	10	40	45	32
translation error	1	0	0	1
translation divergence error	1	0	0	1
alignment error	4	4	5	4.2



University of Edinburgh, University of Groningen

<http://www.ilcc.inf.ed.ac.uk/>, <https://www.rdg.nl/>

Decontextualization: making sentences stand-alone



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

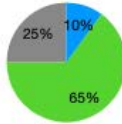
Overview

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.

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)

● Infeasible ● Feasible ● Unnecessary

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

Decontextualization Examples

Document Title: Croatia at the FIFA World Cup
Paragraph: Croatia national football team have appeared in the FIFA world cup on five occasions, since gaining independence in 1991. Before that, from 1930 to 1990 Croatia was part of Yugoslavia. Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.

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 Petricca .
Decontextualized The music video for Shut Up and Dance stars professional dancer Lauren Taft alongside Nicholas Petricca .

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%

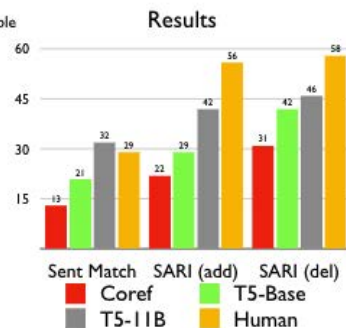
Original In 1850 , the first experimental electric telegraph line was started between Calcutta and Diamond Harbour.
Decontextualized In 1850, the first experimental electric telegraph line in India was started between Calcutta and Diamond Harbour.

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

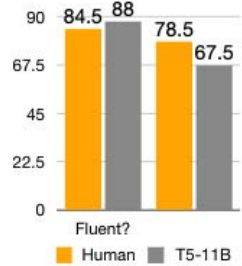
Evaluation Measures
Exact Match (on one of four references, feasible examples only)
SARI-add (added token overlap F1)
SARI-del (deleted token overlap F1)

Models	Coref	SpanBERT model, Untrained baseline
Base line	BART-small	Seq2Seq model, fine-tuned with decontextualized data
	BART-large	
Upper Bounds	Human	Annotator



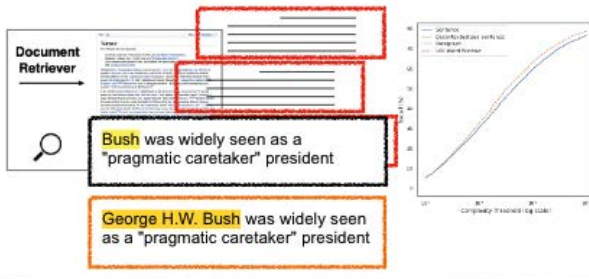
- Well trained seq2seq model performs surprisingly well
- Vanilla evaluation metric is challenging to tell apart good vs. bad decontextualization

Manual Eval



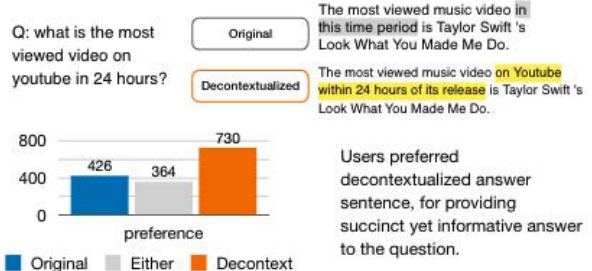
Application 1: Decontextualization as preprocessing

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.



Papers in ML4NLP

Session 8A

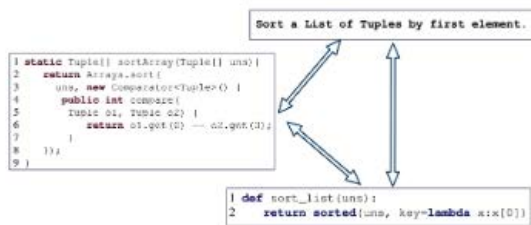
Unified pre-training for program understanding and generation

Unified Pre-training for Program Understanding and Generation

Wasi Uddin Ahmad^{1*}, Saikat Chakraborty^{2*}, Baishakhi Ray², and Kai-Wei Chang¹

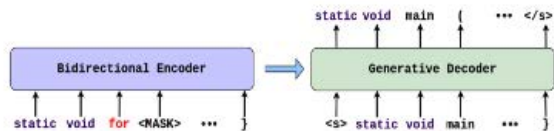
¹University of California, Los Angeles, ²Columbia University

Software Engineering Applications



Our Proposal: PLBART

Pre-train Transformer via denoising autoencoding in program and natural languages.



Denoising Autoencoding

Three noise functions - Token masking, token deletion, token infilling.

PLBART Encoder Input	PLBART Decoder Output
Is 0 the [MASK] Fibonacci [MASK] ? <Eos>	<Eos> Is 0 the first Fibonacci number ?
public static main (String args []) { date = Date (); System . out . (String . format (" Current Date : % ic " ,)) ; } <java>	<java> public static void main (String args []) { Date date = new Date (); System . out . print (String . format (" Current Date : % ic " , date) ; }
def addThreeNumbers (x , y , z) : NEW_LINE INDENT return [MASK] <python>	<python> def addThreeNumbers (x , y , z) : NEW_LINE INDENT return x + y + z

Evaluation: Code Summarization

Dataset: CodeSearchNet
Evaluation Metric: BLEU-4

Methods	Ruby	Javascript	Go	Python	Java	PHP	Overall
Seq2Seq	9.64	10.21	13.98	15.93	15.09	21.08	14.32
Transformer	11.18	11.59	16.38	15.81	16.26	22.12	15.56
RoBERTa	11.17	11.90	17.72	18.14	16.47	24.02	16.57
CodeBERT	12.16	14.90	18.07	19.06	17.65	25.16	17.83
PLBART	14.11	15.56	18.91	19.30	18.45	23.58	18.32

Evaluation: Code Translation

Dataset: CodeXGLUE
Metric: BLEU, EM (exact match), CodeBLEU

Methods	Java to C#			C# to Java		
	BLEU	EM	CodeBLEU	BLEU	EM	CodeBLEU
Transformer	55.84	33.00	63.74	50.47	37.90	61.59
RoBERTa (code)	77.46	56.10	83.07	71.99	57.90	80.18
CodeBERT	79.92	59.00	85.10	72.14	58.80	79.41
GraphCodeBERT	80.58	59.40	-	72.64	58.80	-
PLBART	83.02	64.60	87.92	78.35	65.00	85.27

Qualitative Example: Code Translation

Input Code: C#	Generated Code: Java
<pre> 1 public int GetCells() { 2 int size = 0; 3 foreach (char c in cells.Keys) { 4 int n = 0; 5 if (c.Length > 0 c.IsDigit()) { 6 size++; 7 } 8 } 9 return size; 10 } </pre>	<pre> 1 public int GetCells() { 2 Iterator<Character> i = 3 cells.keySet().iterator(); 4 int size = 0; 5 for (i.hasNext(); i.hasNext(); i.next()) { 6 Character c = i.next(); 7 if (c.length() > 0 c.isDigit()) { 8 size++; 9 } 10 } 11 } 12 return size; 13 } </pre>

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 fine-tuning time to only use the internal representations of the model.

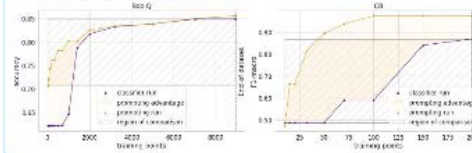
In contrast, a prompted model is used as a token predictor. It is presented with the input of the task and a *prompt*: a short description of the task that needs to be completed by a masked token. That token is then mapped into the desired class output by a *verbalizer*. (for example, Yes for 1 and No 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 (of Training Points)					
	MNLI	BoolQ	CB	CDQA	MultiRC ²	RTE
<i>P vs H</i>	3506 ± 536	752 ± 46	90 ± 2	258 ± 242	384 ± 178	292 ± 34
<i>P vs N</i>	150 ± 252	229 ± 81	78 ± 2	-	74 ± 56	80 ± 68
<i>N vs H</i>	3355 ± 612	453 ± 90	12 ± 1	-	309 ± 320	-22 ± 62

4/ Zero-shot vs adaptation

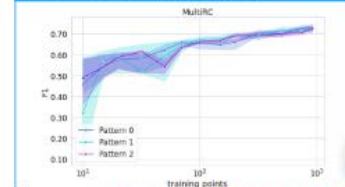
In order to study the zero-shot vs. adaptive nature of prompts, we introduce *null verbalizers*: models whose verbalizers are replaced with first names, so that zero-shot capability is at random chance. We compute which part of the data advantage is due to the zero-shot capability (*prompt vs null*)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

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

	MultiRC ²	RTE	WIC
	384 ± 178	252 ± 34	424 ± 71
	74 ± 56	404 ± 68	-354 ± 166
	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 to 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 | ar@protonmail.com
[†]University of Massachusetts Lowell | {kovaleva, arum}@cs.uml.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 (Hevrit et al. 2019)
- BERT's MLM prefers correct to incorrect verb forms (Goldberg 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)
- BERT can't reason over the facts it "knows" (Forbes et al. 2019)
- BERT's knowledge sometimes comes from stereotypical associations (Pfeiffer et al. 2019)

3. WHAT DOES PROBING SHOW? 😊

- Probing classifiers can extract a lot of information from BERT embeddings about part of speech, syntactic chunks and roles, etc. (Liu et al. 2019)
- Words sharing syntactic subtypes have larger impact on each other in MLM (Wu et al. 2020)
- Issues with probing:
 - different probing methods may lead to complementary or even contradictory conclusions (Wortschke et al. 2019)
 - the fact that an 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 prediction with other attention (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 (Pfeiffer 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 (Shubell et al. 2019)
- Most heads & layers can be pruned without much impact on performance (Michel et al. 2019; Kovaleva et al. 2019; Volta 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:
 - improvements to the training regime (Liu et al. 2019, ...)
 - tweaking the pre-training (Joshi et al. 2020, ...)
 - tweaking the fine-tuning (Arase 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 relevant BERT weights are frozen (McCoy et al. 2019)
- Some 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 NLI (McCoy et al. 2019; Zellers et al. 2019; Jin et al. 2020)
 - BERT learns shallow heuristics in QNLI (Rogers et al. 2020; Sugriva et al. 2020)
 - Probably also everywhere else
- BERT works pretty well even without pre-training! (Kovaleva et al. 2019)
- Shallow heuristics can be used to reconstruct the model (Kishita 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? (amnesic probing (Etazar et al. 2020), pruning to interpret (Volta et al. 2019; Pfeiffer 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
 - uniformly 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 1^{st} 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.



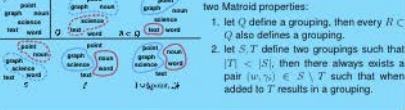
Entropy-based Diverse Grouping

- 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 $\frac{1-\epsilon}{2}$ -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

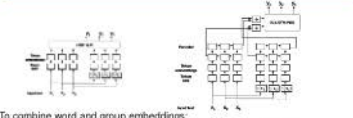


Submodular Function

- 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 / c_w : word w / group relative frequency
 - the entropy gain only depend upon partial entropies of group index i

$$-(c_{\gamma_i} + F_w) \log(c_{\gamma_i} + F_w) + c_{\gamma_i} \log(c_{\gamma_i}) - (c_{\gamma_j} + F_w) \log(c_{\gamma_j} + F_w) + c_{\gamma_j} \log(c_{\gamma_j})$$
 - thus $c_{\gamma_i} < c_{\gamma_j}$ where c_{γ_i} and c_{γ_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

Machine Translation (in BLEU%)	IWSLT17 EN-FR	Language Modeling (in F1%)	IWSLT17 EN
Baseline (ConvS2S)	17.6	Baseline (RLM)	22.65
Random Grouping	23.3	Entropy-based Grouping	21.99 (-3.76%)
Poisson-based Random Grouping	22.4		
Distance-based Grouping	23.6		
Entropy-based Grouping	24.1 (+6.9%)		

Machine Translation (in BLEU%)	IWSLT17	MTNT'16
Baseline (ConvS2S)	19.4	22.6
Entropy-based Grouping	21.0	24.0
	19.2	23.9 (+18.8%)

Part-of-Speech Tagging	Brown Corpus EN		
	Loss	Accuracy(%)	Error Rate(%)
Baseline (S3T)	5.16	93.61	1.39
Random Grouping	5.32	93.07	1.31
Poisson-based Random Grouping	5.50	92.27	1.39
Entropy-based Grouping	5.48	92.65	1.34 (-3.60%)

Acknowledgements

- National Science Foundation (NSF) Award No. 1747728
- National Science Foundation of China (NSFC) Award No. 61672524
- 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

KEY TAKEAWAYS

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

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

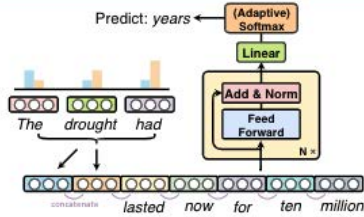
ABLATION

Model Config	# params	Val. Perplexity
Transformer	148M	25.0
NPLM-old	32M	216.0
NPLM-old (large)	221M	128.2
NPLM 1L	123M	52.8
NPLM 4L	128M	38.3
NPLM 16L	148M	31.7
- 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

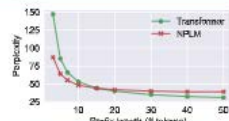
- Three word-level benchmarks (WIKITEXT-2, WIKITEXT-103, LAMBADA) and one character-level benchmark (ENWIK8).

Model Config	WIKITEXT-2	WIKITEXT-103	LAMBADA	ENWIK8
NPLM	120.5	31.7	44.8	1.63
Transformer	117.6	25.0	42.1	1.14
Transformer-C	113.1	24.1	42.0	1.14
Transformer-N	110.8	24.1	41.8	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 (↓)	Control (↓)
NPLM	0.4	30.46
Transformer	30.60	35.84
Transformer-N	32.51	37.06
Transformer-C	32.23	37.34

Token type	CF (↓)	LF (↓)	Ent. (↓)
Transformer	38.94	29.47	32.26
Transformer-N	42.33	30.14	33.95
Transformer-C	42.65	31.58	35.03

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^{1*}, Aaron Jaech¹, Xin Li¹, Matt Gormley¹, Jason Eisner²

¹Johns Hopkins University
²Facebook 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 $p(x) \propto \tilde{p}(x)$ where $\tilde{p}(x)$ 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 $O(\text{poly}(|x_i|))$.

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 θ_n to compute $p(x) = \prod_i p(x_i | x_{<i})$. They provide a conceptual upper bound to the just-train-a-slightly-larger-model paradigm 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 $p(x|y) > 0$ if y is a valid continuation of prompt x , even if that property can be checked in polytime.

P/poly:

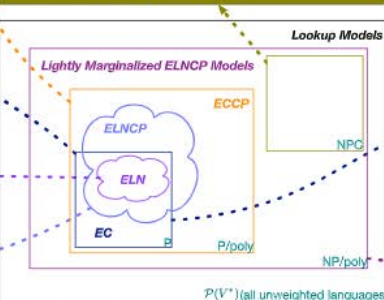
P with the help of poly-sized advice strings that can come from an oracle. P/poly is therefore more powerful than P — they can model undecidable problems due to the oracle access!

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 support of 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

$$p(x \# y) = p(x) \cdot p(y | x)$$

where $p(x \# y) \neq 0$ iff y is a solution to 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/polysize (i.e. ELN/ELNCP)! The other sequence order does fine under autoregressive models (if x is in NP):

$$p(y \# x) = p(y) \cdot p(x | y \#)$$

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:

$$p(x \# y) = p(x) \cdot p(y | x) = p(x) \cdot \sum_{z} p(y | x, z) \cdot p(z | x)$$

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

$$p(x \# y) = \sum_{z} p(x) \cdot p(y | x, z) \cdot p(z | x)$$

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: kNNLM, adaptive semi-parametric language models, ...
Downside: Need a vast database of observed or precomputed answers.

Model family	Efficiently computable?	Efficiently locally normalized?	Efficiently locally normalizable with compact parameters?	Support can be ...
EBMs/EBMs+Autoregressive models (ELNCP)	✓	✓	✓	arbitrary
Autoregressive models (ELN)	✓	✓	✓	arbitrary
Lightly marginalized ELNCP	✓	✓	✓	arbitrary
Lookup models (ELNCP)	✓	✓	✓	arbitrary

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

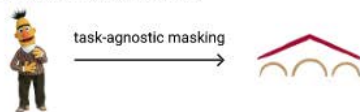
On the Inductive Bias of Masked Language Modeling: From Statistical to Syntactic Dependencies

Tianyi Zhang, Tatsunori B. Hashimoto



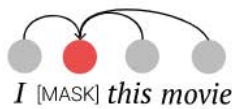
code release@github:
tatsu-lab/mlm_inductive_bias

1 Problem Statement



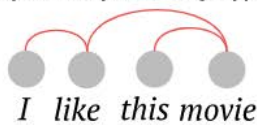
Question: why can MLM capture linguistic structure and transfer to new tasks?

2.1 Cloze Reduction Hypothesis



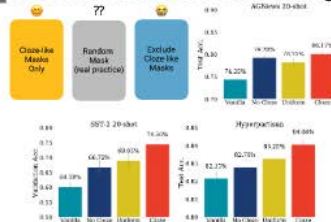
Cloze-like masking can provide indirect supervision but in practice we apply random masking. We quantify the importance of cloze-like masking through controlled experiments.

2.2 Dependency Learning Hypothesis



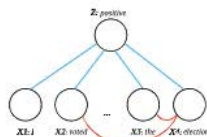
We hypothesize that generic masks can help learn the statistical dependencies among words and these dependencies are related to syntactic structures

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 $\Sigma_{\mathbf{X}\mathbf{X}}$ is full rank,

$$\mathbf{x}_{\text{mask},i} = \beta_{2SL_{i,j}} \mathbf{X}_{i,j} + O(\|\Sigma_{\mathbf{X}\mathbf{X} \setminus \{i,j}\}\|_2)$$

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 $H(\mathbf{Z} | \mathbf{X}_{\setminus \{i,j\}})$.

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

MLM objective directly ensures good approximation of cond. MI.

Proposition 4. Let

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

be an estimator constructed by the model distribution p_{θ} . Then we can show

$$|I_{ij} - \hat{I}_{ij}| \leq \mathbb{E}_{x_i} D_{KL}(p(x_i | X_{\setminus \{i,j\}}) \| p(x_i | X_{\setminus \{i,j\}}, x_j))$$

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 ± 0.73
LINEARCHAIN	54.13
Klein and Manning (2004)	55.91 ± 0.68
PMI	33.94
CONDITIONAL PMI	52.44 ± 0.19
CONDITIONAL MI	58.74 ± 0.22

Table 1: Unlabeled Undirected Attachment Score on WSJ10 test split (section 23). Error bars show standard deviation across three random seeds.