NAACL 2021 papers

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

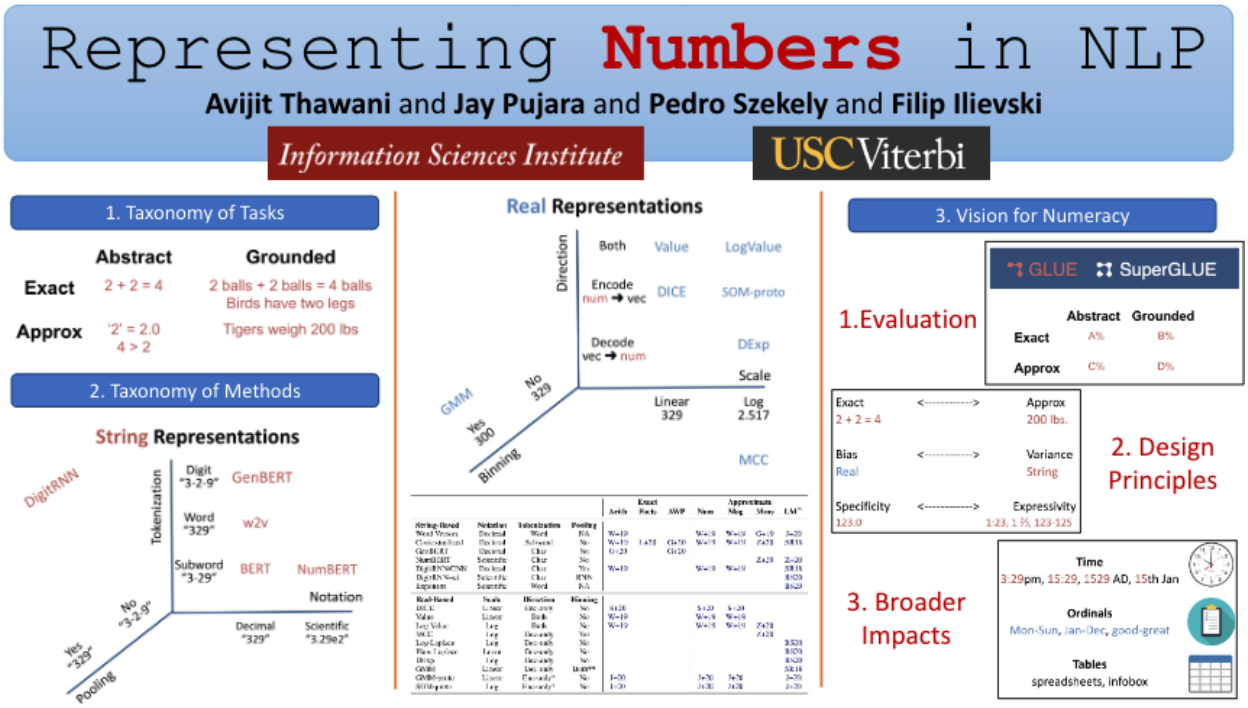
Tutorials

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

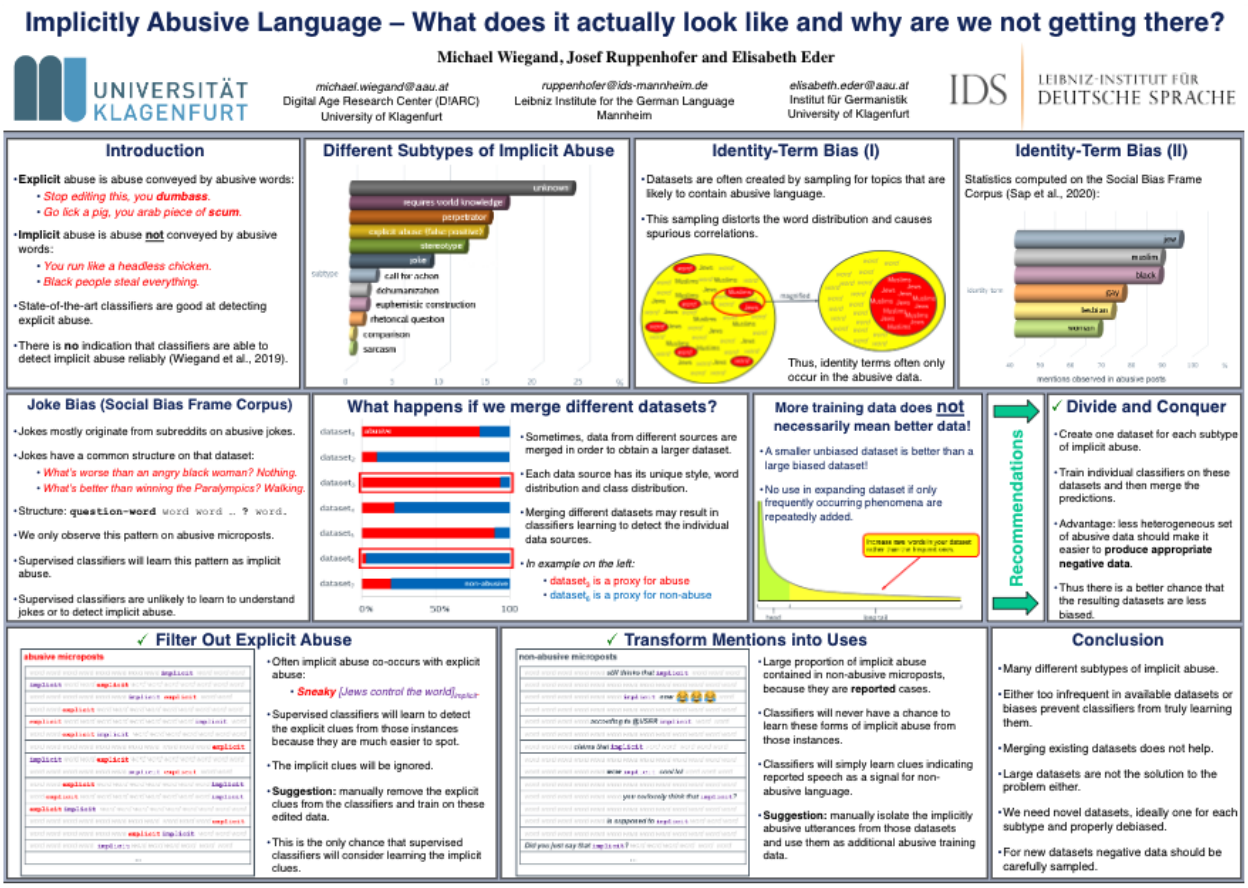
Special Papers

Session 2D (New Challenges)

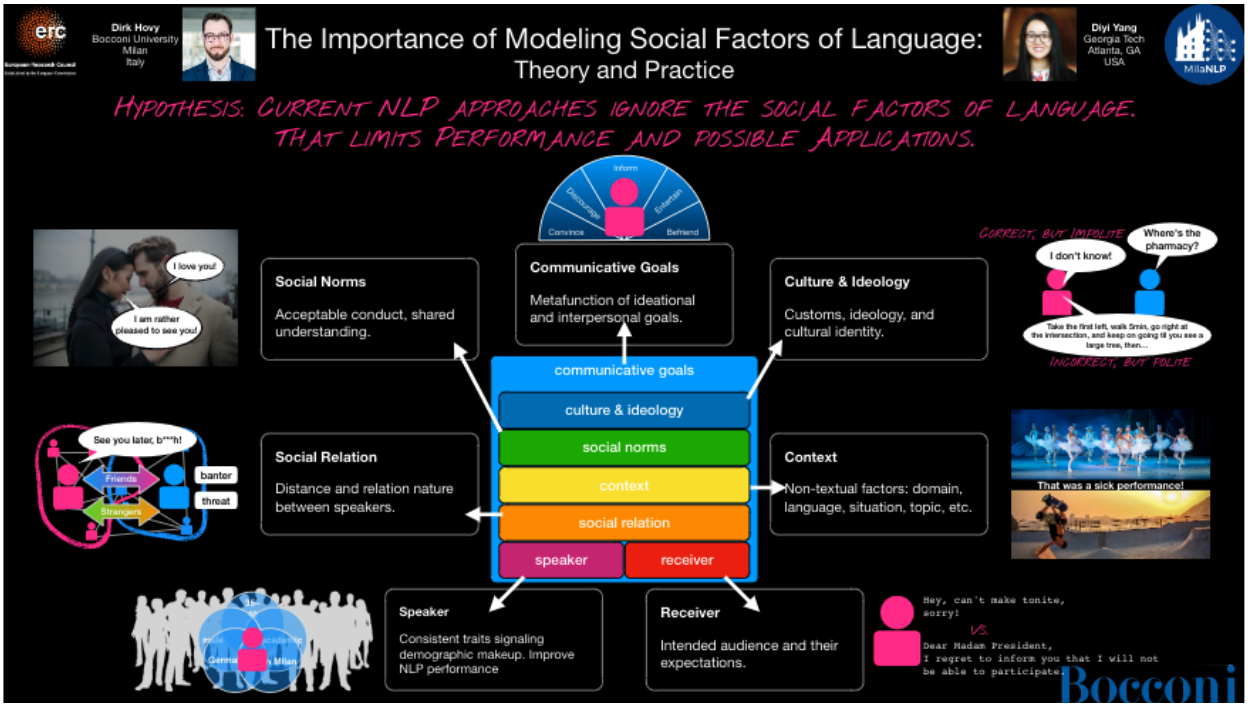
Representing numbers in NLP: a survey and a vision



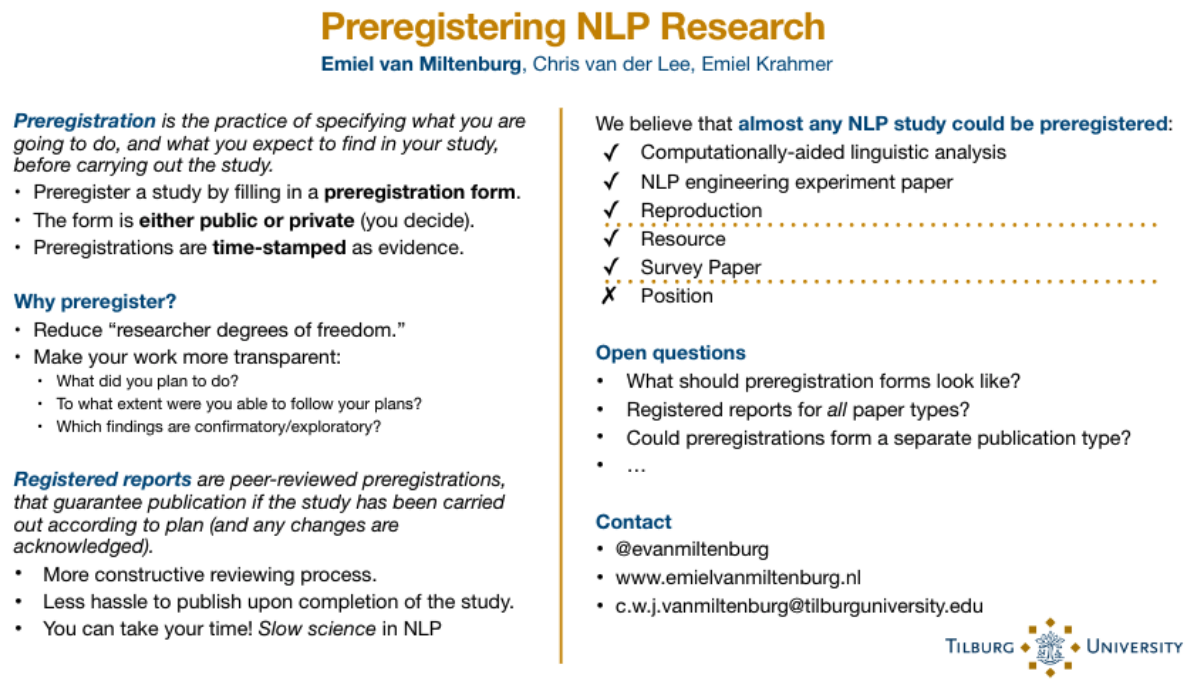
Implicitly abusive language — what does it actually look like and why are we not getting there?



The importance of modeling social factors of language



Preregistering NLP research



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

Stanford NLP

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.

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

TSE

P(**are**) > P(is), so the score is 1.0

Likely Behavior

P(correctly conjugated verb) = P(**are**) + P(**exist**) = 0.7

TSE overestimates this because it assigns a binary score to each verb.

Systematicity

P(**are**) > P(is)
P(**exist**) < P(exists) ← TSE misses this because it's not in the minimal pair dataset.

Our Metrics

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

Likely Behavior: **Model Weighted Syntactic Evaluation (MW)**

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

E.g.

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

Systematicity: **Equally Weighted Syntactic Evaluation (EW)**

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

E.g.

$$\frac{1}{2}\left(\mathbb{1}\left[P_M(\textbf{are}) > P_M(\text{is})\right] + \mathbb{1}\left[P_M(\textbf{exist}) > P_M(\text{exists})\right]\right) = 0.5$$

Evaluations

Minimal Pairs Datasets
- Marvin and Linzen[2]
- BLiMP[3]

Lemmas: ~3500
- COCA[4]
- Penn Treebank[5]
- Giant Verb List[6]

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

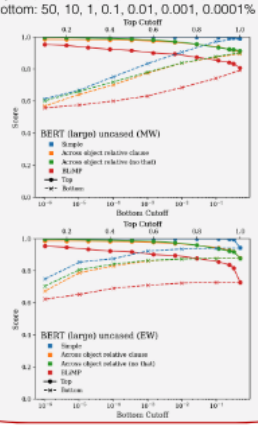
Templates	BERT cased			BERT uncased			RoBERTa			
	MW	EW	TSE	MW	EW	TSE	MW	EW	TSE	MW
Simple	0.99	0.94	**1.00**	0.98	0.90	**1.00**	0.98	0.93	**1.00**	0.90
In a sentential complement	0.92	0.67	**0.89**	0.92	0.60	**0.86**	0.92	0.67	**0.88**	0.96
VP coordination	0.91	0.89	**0.90**	0.93	**0.90**	**0.90**	0.93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

Qualitative Examples

The senators that the skater ___ are young	meets	encounters	sees	meet	encounter	see	glide	faces	sees	met
	0.20	0.057	0.057	0.048	0.023	0.029	0.018	0.013	0.012	0.010
The pilots that the executive ___ are tall	pilots	controls	employs	trains	uses	includes	has	orders	commands	features
	0.088	0.059	0.025	0.023	0.022	0.009	0.017	0.015	0.014	0.013

Likely Verbs

Why are our scores low?
We look at the top and bottom p% of models' distributions:
- Top: 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 97, 100%
- Bottom: 50, 10, 1, 0.1, 0.01, 0.001, 0.0001%



Conclusion

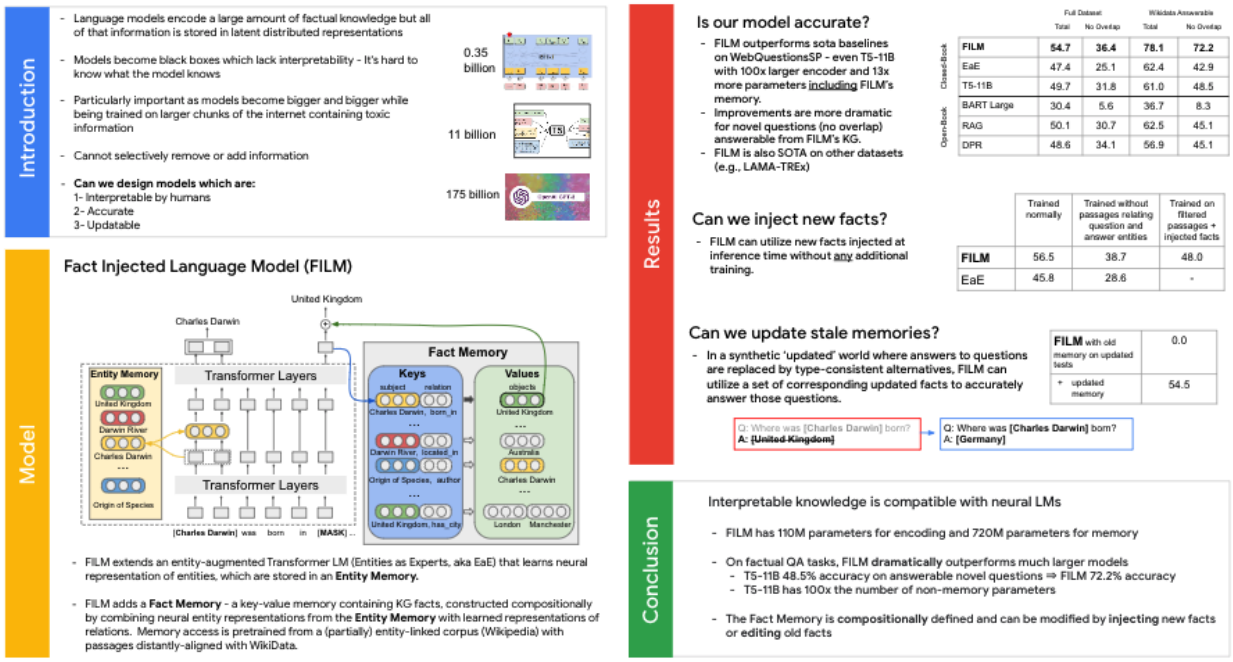
- We refine **TSE** to measure **likely behavior** (with **MW**) and **systematicity** (with **EW**) of language models
- We find models more often correctly conjugate verbs they deem likely

[1] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. TACL.
[2] Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. EMNLP.
[3] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. TACL.
[4] Mark Davies. 2008. The corpus of contemporary American English.
[5] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. CL.
[6] Pattern Based Writing: Quick & Easy Essay. 2019. Giant verb list. 3,200 verbs plus spelling rules and irregular 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, liviobs, wcohen}@google.com, haitians@cs.cmu.edu
Google Research



		Full Dataset: Total	Full Dataset: No Overlap	Wikidata Answerable: Total	Wikidata Answerable: No Overlap
Closed-Book	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
Open-Book	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

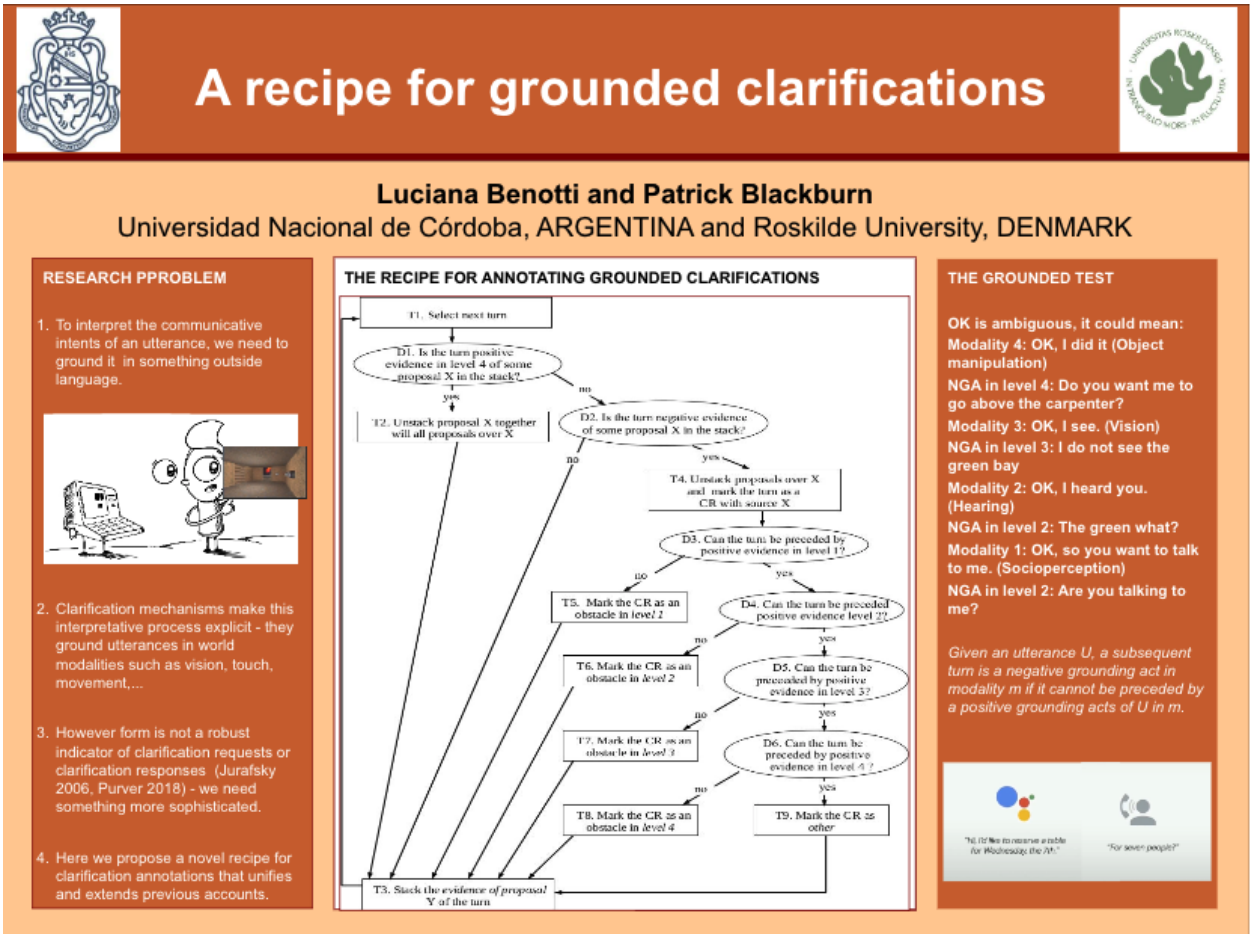
	Trained normally	Trained without passages relating question and answer entities	Trained on filtered passages + injected facts
FILM	56.5	38.7	48.0
EaE	45.8	28.6	-

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

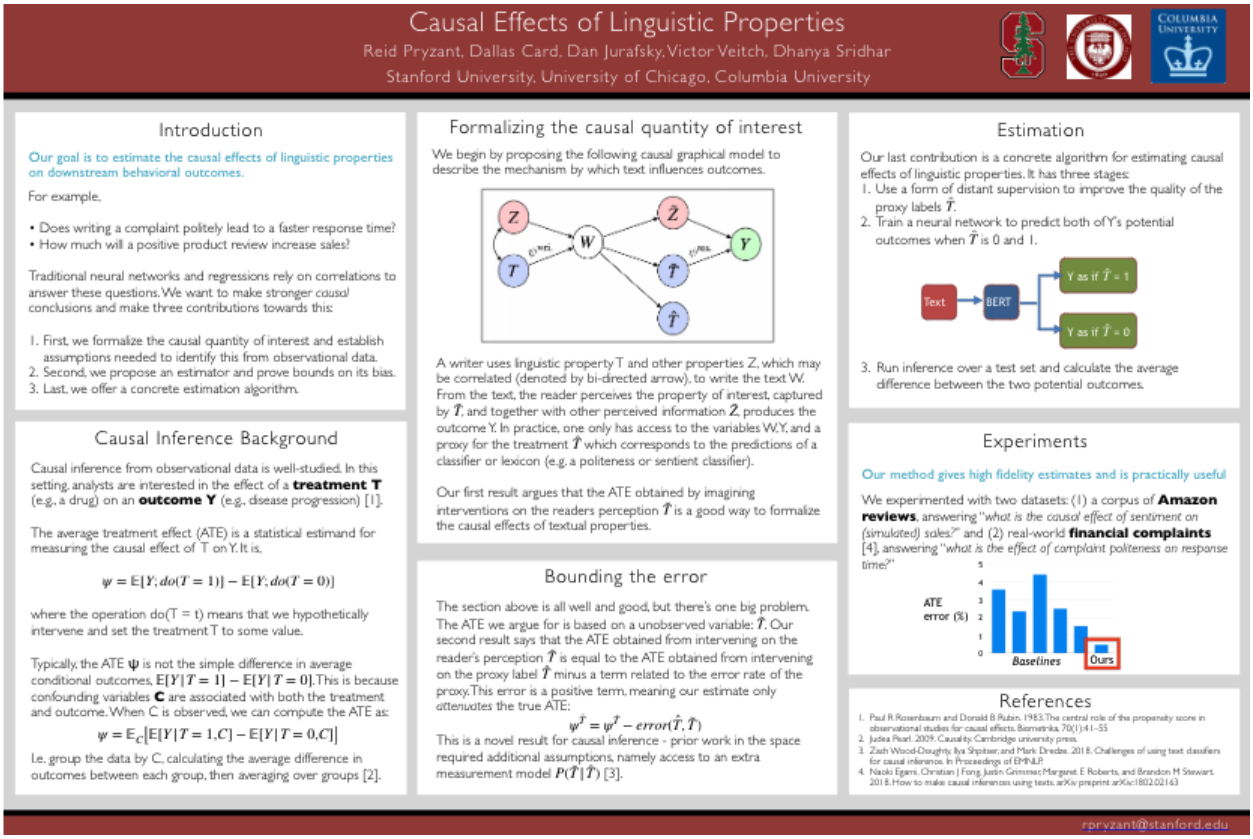
*equal contribution, work done at Google

Session 11E (Special Theme)

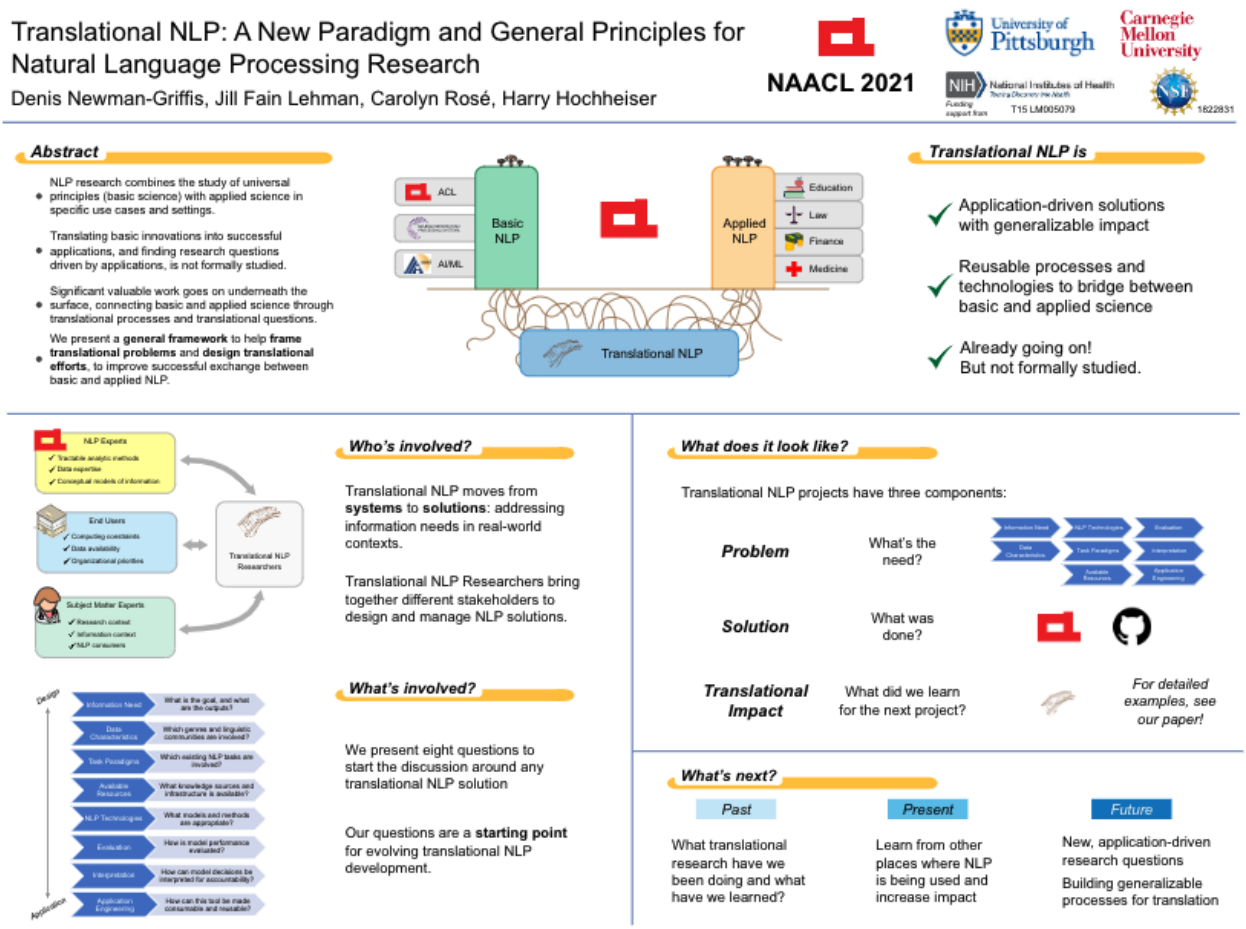
A recipe for annotating grounded clarifications



Causal effects of linguistic properties



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



Papers in interpretability

Session 1B

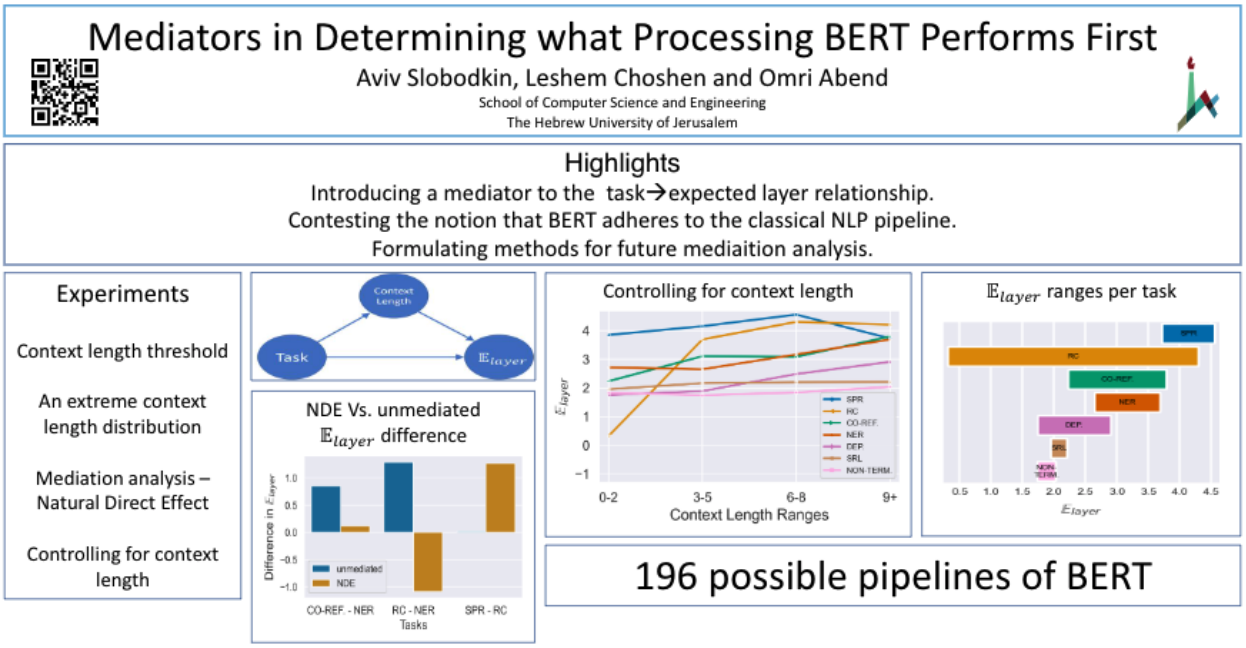
Concealed data poisoning attacks on NLP models

- 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

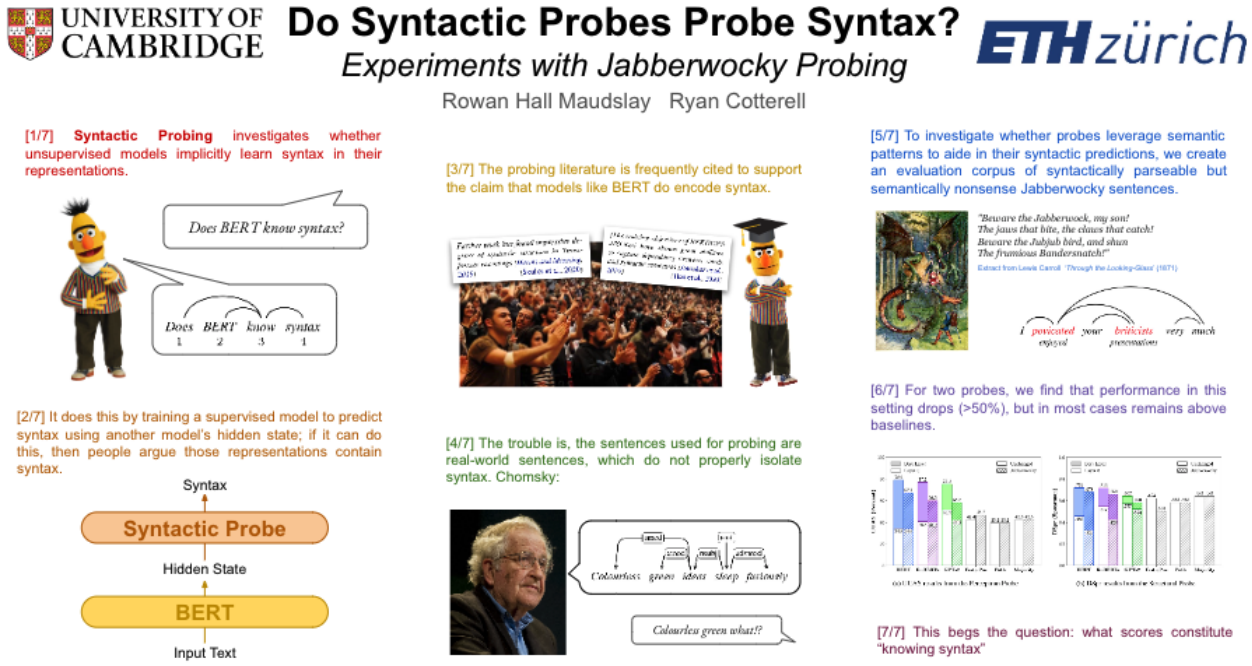


Automatic generation of contrast sets from scene graphs: probing the compositional consistency of GQA

- 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



- 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 o the spatial domain and of the natural language instructions relative to it, or whether they merely overfit spurious signals in the dataset.

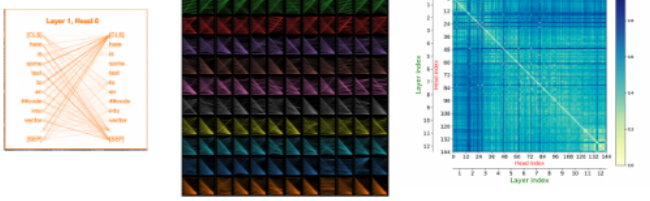
On attention redundancy: a comprehensive study

On Attention Redundancy: A Comprehensive Study

NAACL 2021

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

Motivation



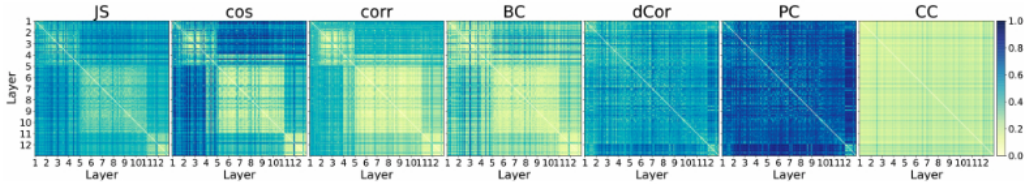
Attention head | L-layer-H-head | Redundancy matrix (What)

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

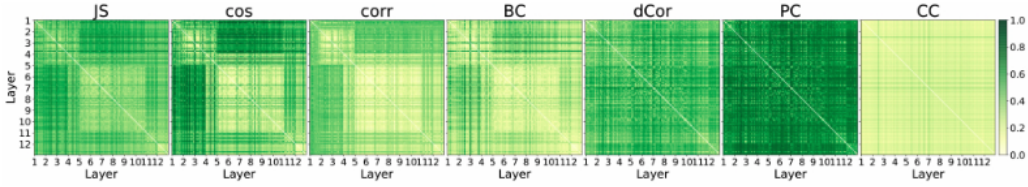
Experimental Results

How & Where



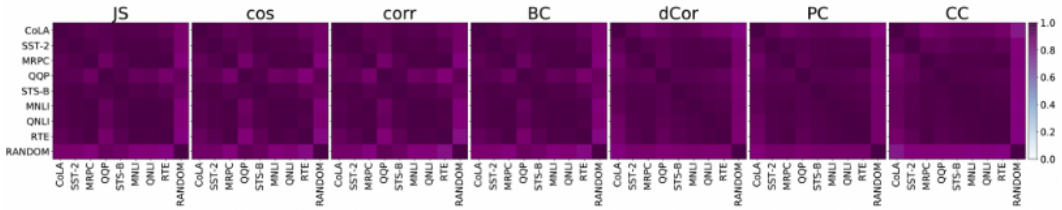
How and Where: token-based (left 4) and sentence-based distances (right 3) (pre-trained BERT-base on CoLA)

When



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

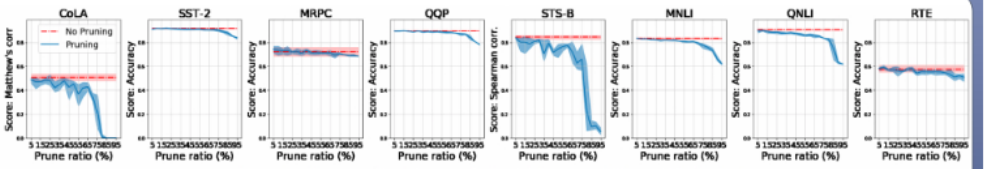
Who



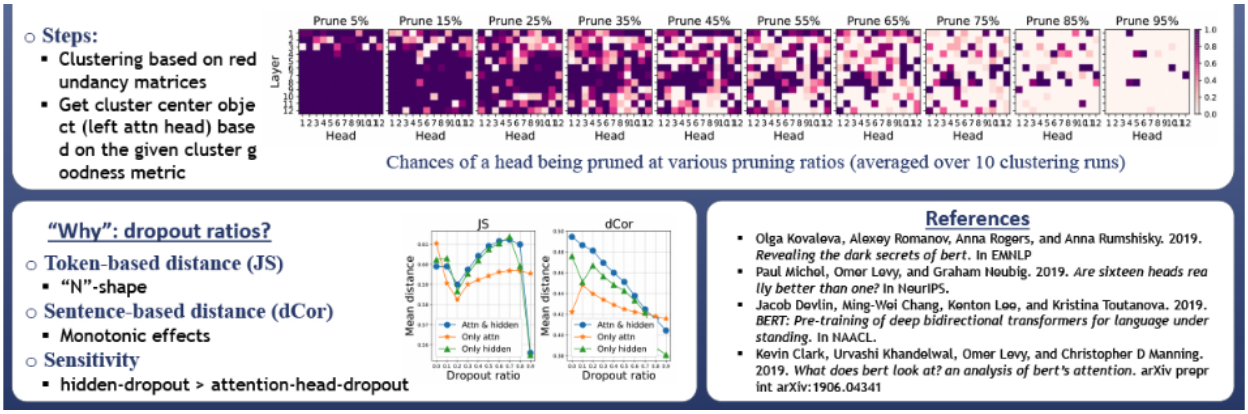
Who: Task-agnostic: Correlations of redundancy matrices of each pair of tasks (pre-trained BERT-base on GLUE tasks)

Zero-shot head pruning:

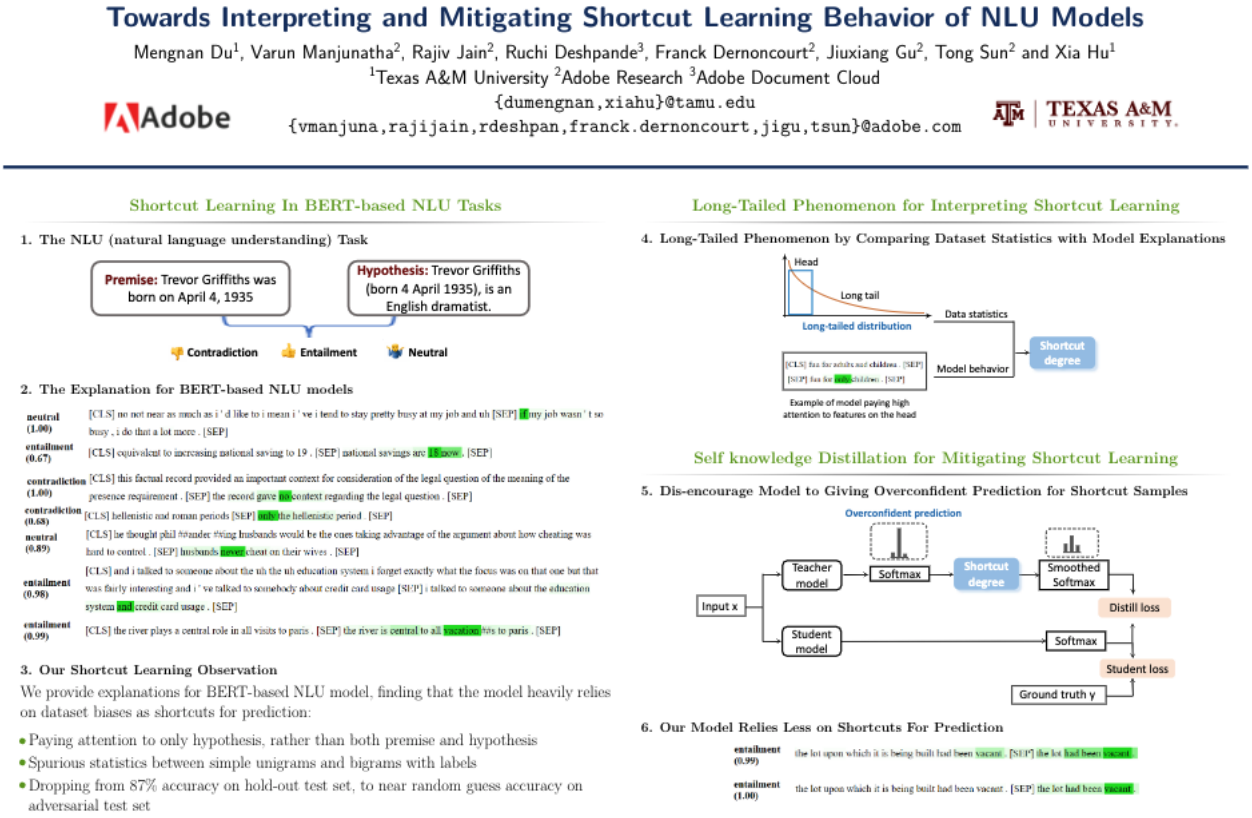
- **Input:**
  - Redundancy matrices
  - Model: pretrained BERT
  - Data: randomly generated token sequences (no data of downstream tasks is required)



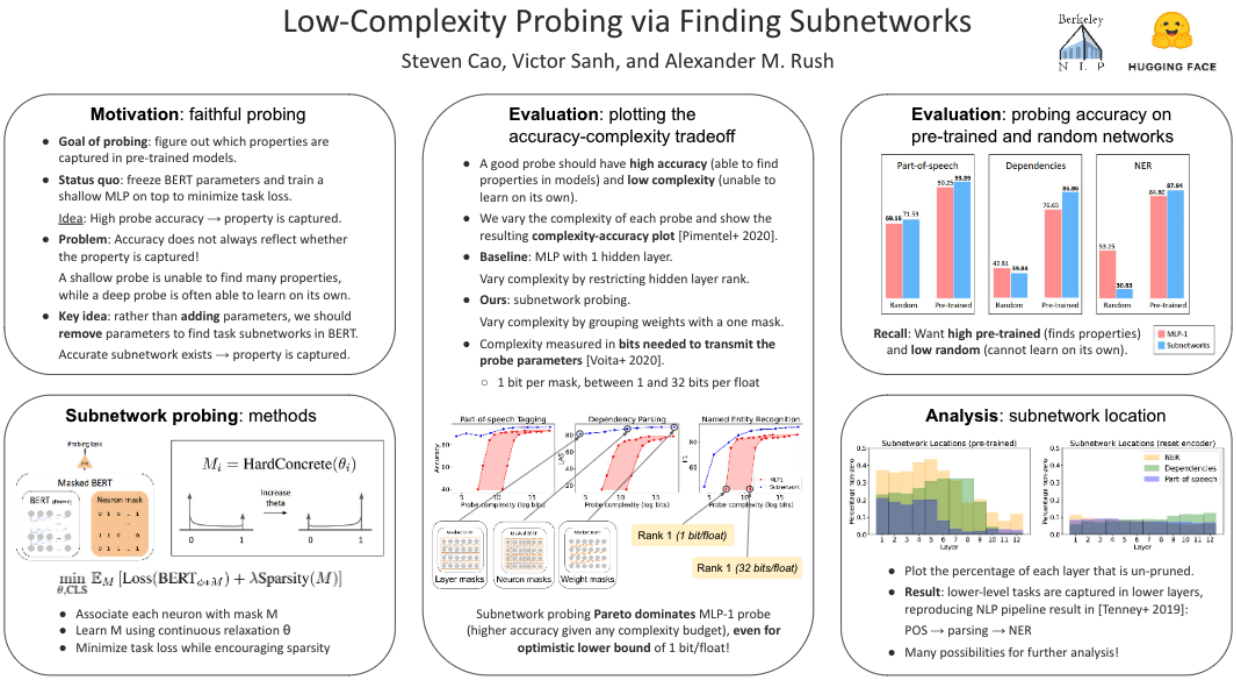
Fine-tuning results of pruned BERT-base model on GLUE tasks
(75 – 85% heads can be pruned to preserve performances for most tasks)



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

- 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.

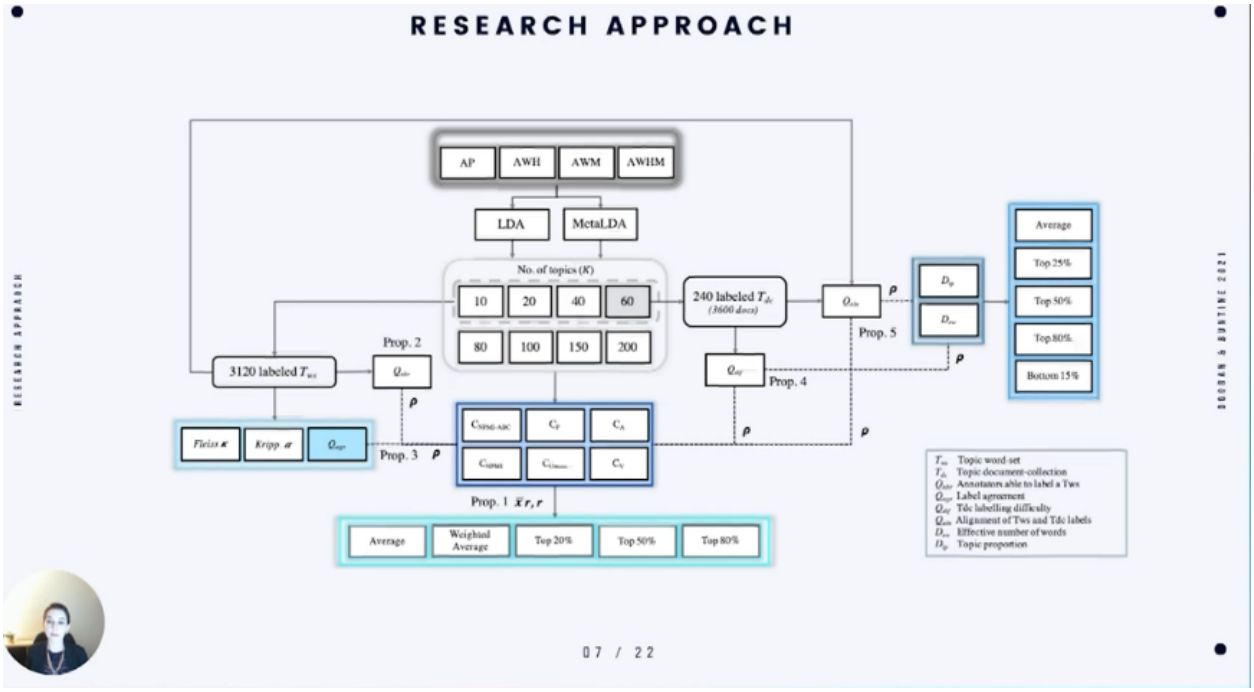
- Qagr: agreement on the labels given to a topic between the four SME as a percentage.

4. An interpretable topic is one where the document collection is easily labelled.

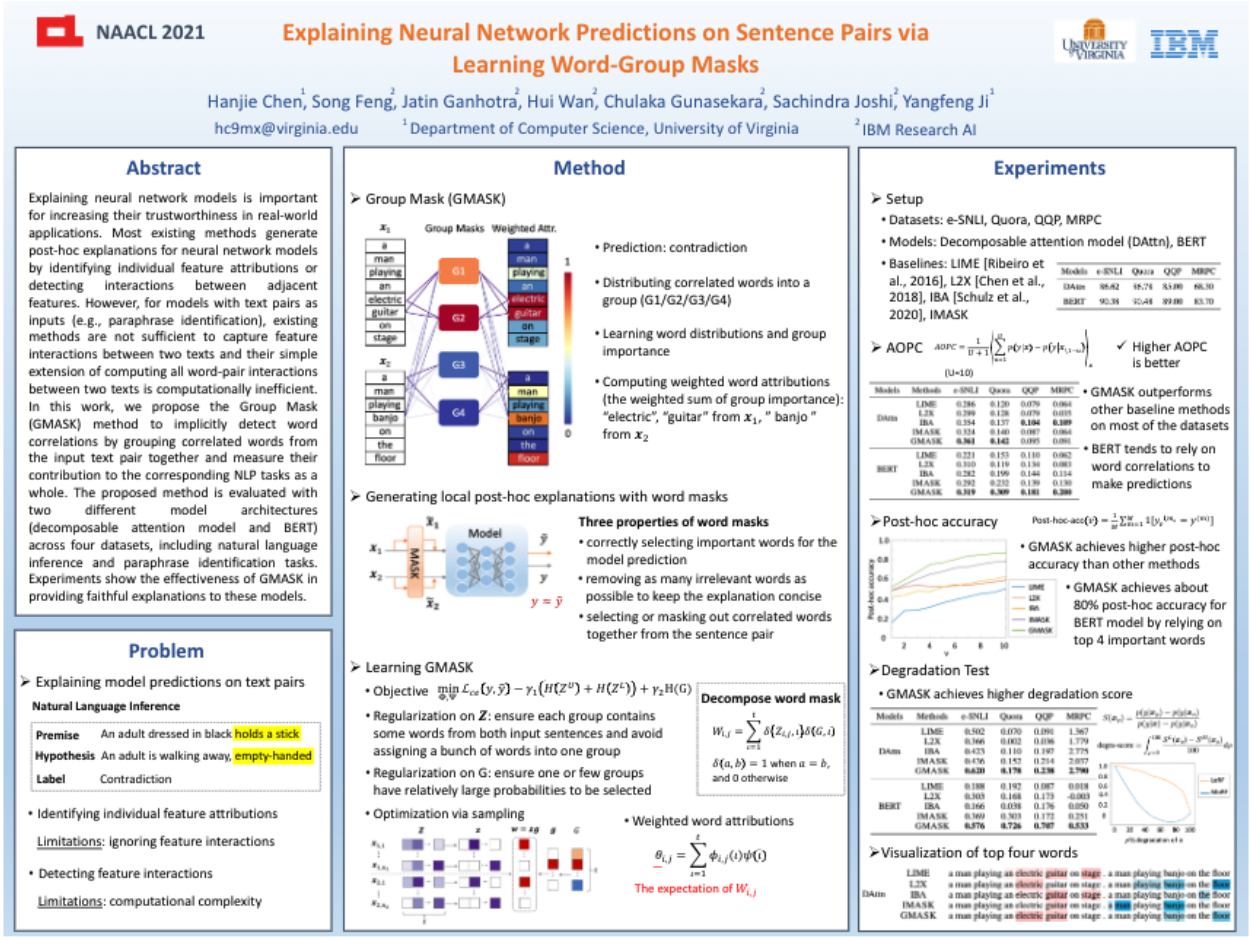
    - Qdiff: Labeling difficulty.
    - Qdiff and coherence, Qdiff and Dew: some relationships. No relationship between Qdiff and Dtp.

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



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)

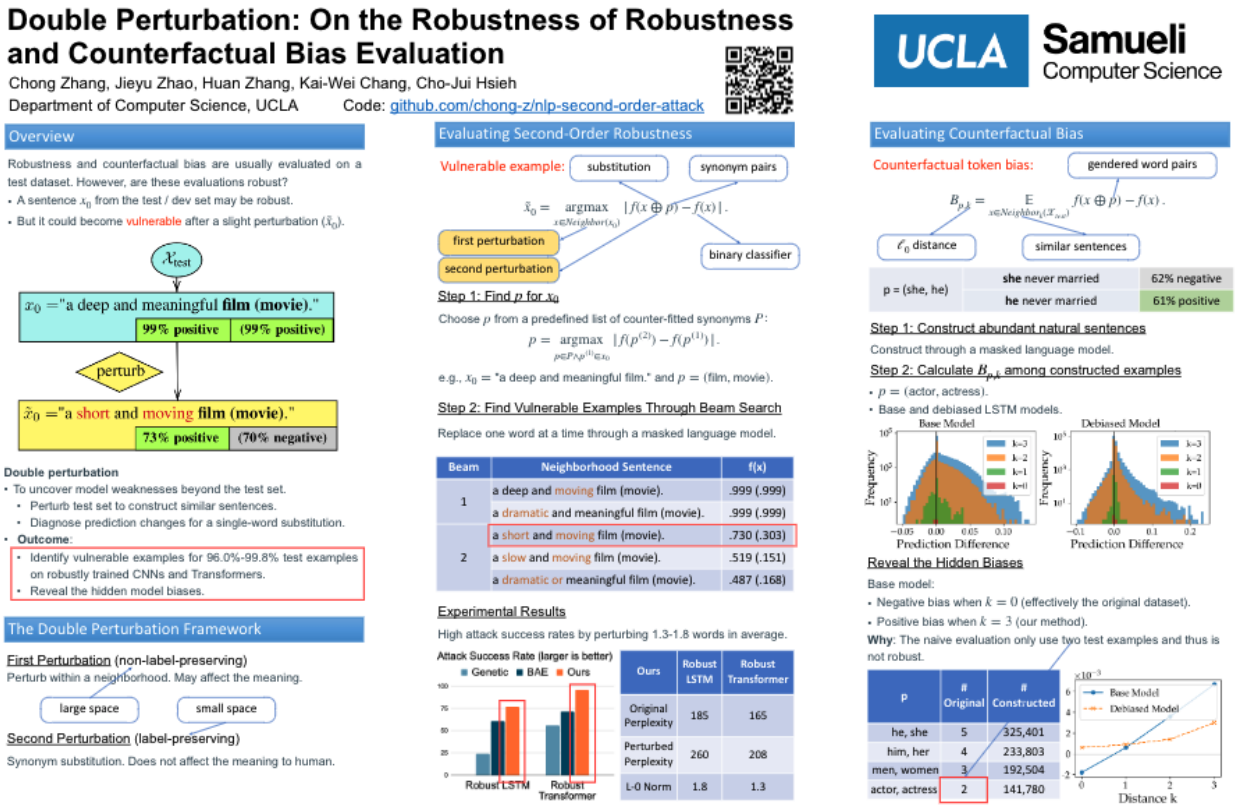
Findings:

- 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



- 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=7ZL84tVlHZN
- 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-

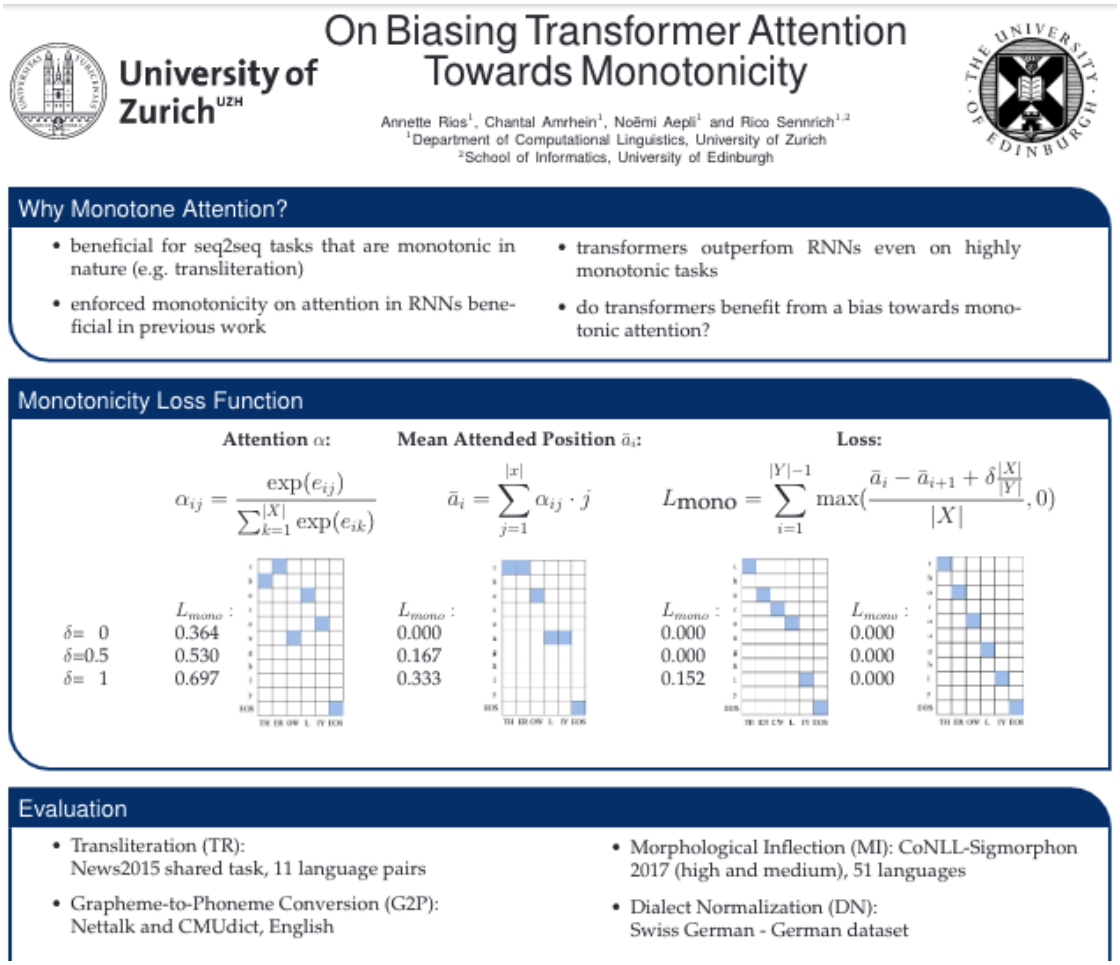
Paper.pdf

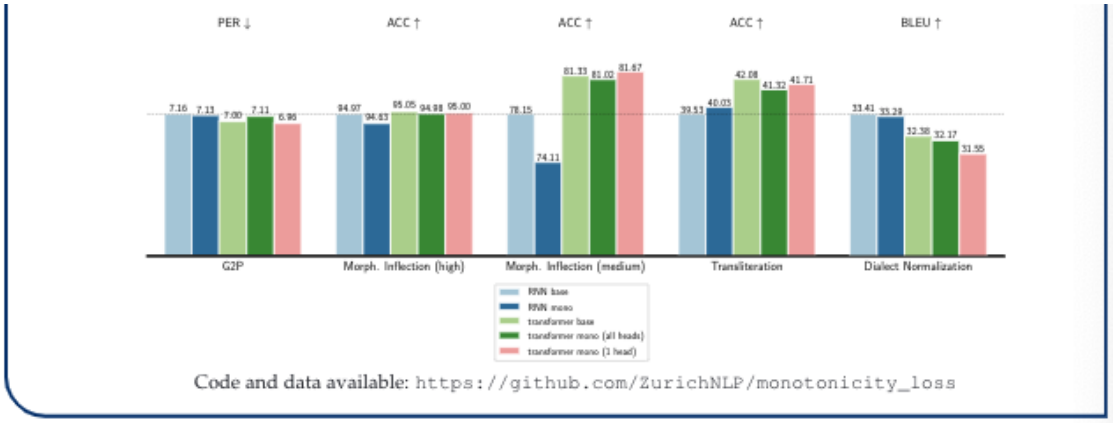
- Towards hierarchical importance attribution: explaining compositional semantics for neural sequence models https://arxiv.org/abs/1911.06194
- Probing with Shapley-value-based explanations as feature importance measures http://proceedings.mlr.press/v119/kumar20e/kumar20e.pdf

Papers in linguistic theory, psycholinguistics

Session 12E

On biasing Transformer attention towards monotonicity





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

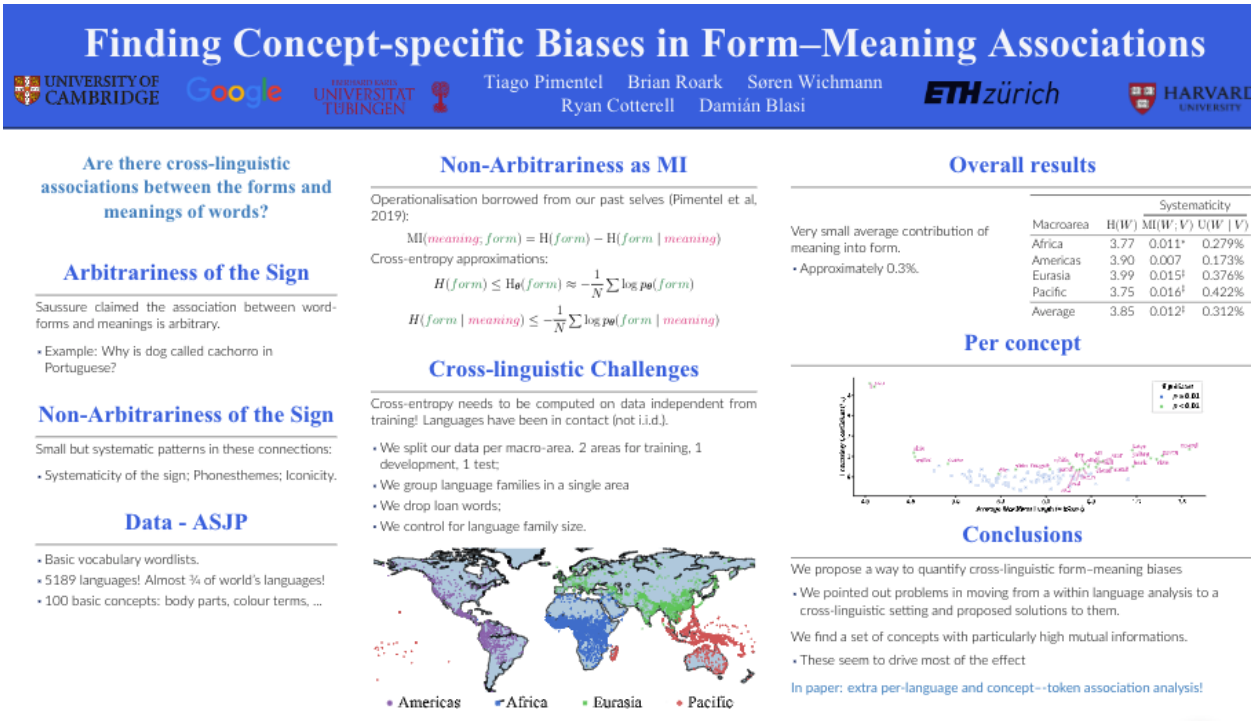
Takeaway

- monotonicity in attention increased across all tasks and datasets
- mixed results: improvements on some tasks
- transformers: loss on all heads impairs ability to learn specialized functions
- loss on subset of heads: beneficial on some tasks

future work: explore usefulness of loss

- where alignment is harder to learn
- with more complex training schedule

Finding concept-specific biases in form-meaning associations



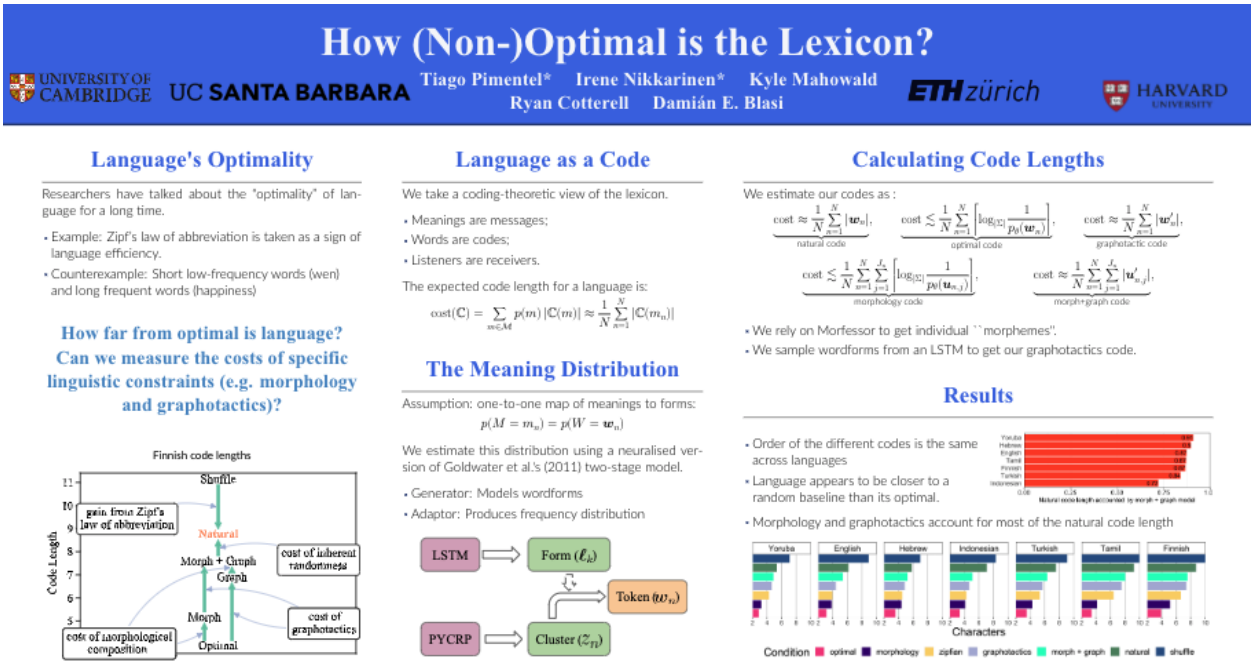
Ab Antiquo: neuro proto-language reconstruction

- 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 Wikitionary.
- Synthetic evaluation dataset
- Analysis of learned representation reveals the learning of phonologically meaningful representations without direct supervision.

How (non-)optimal is the lexicon?



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.

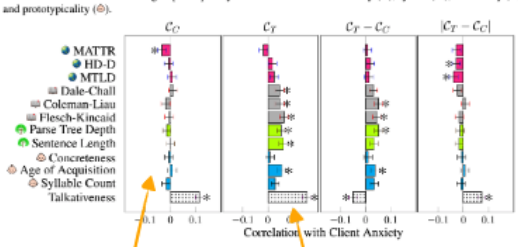
	Dataset	
	Exploratory	Confirmatory
Messages	2.6 million	0.7 million
Survey responses	24,287	6,150
Clients	5,736	1,434
Therapists	1,608	889
†Survey responses / client	4.23	4.29
†Client text (words) / survey	1259	1295
†Therapist text (words) / survey	796	804
Median survey score (0-21)	8	8
Median time between 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$99.9% confidence intervals) on the exploratory dataset for language complexity of clients ($\mathcal{C}_C$), therapists ($\mathcal{C}_T$), therapist and client difference ($\mathcal{C}_T - \mathcal{C}_C$), and absolute therapist and client difference ($|\mathcal{C}_T - \mathcal{C}_C|$). Each complexity measure was entered into its own linear mixed model. We group complexity measures into lexical diversity (●), syntax (●), readability (▤), and prototypicality (●).



When clients were more anxious, their linguistic complexity dropped.

When clients were more anxious, therapist linguistic complexity increased.

Variation in Complexity Loss

Table 3: z indicates how much individuals varied linguistic complexity among their own messages compared with a random sample from the population. We show average z 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)			
	z^{σ}_{T}	z^{σ}_{C}	$z^{\sigma}_{T} \neq z^{\sigma}_{C}$? t	p	z^{Δ}_{T}	z^{Δ}_{C}	$z^{\Delta}_{T} \neq z^{\Delta}_{C}$? t	p
MATTR	-0.36	-0.33	-0.87	0.3842	-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
MTLD	-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	0.0664
Flesch-Kincaid	-0.46	-0.93	15.68*	<0.0001	-0.47	-0.76	11.35*	<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	13.54*	<0.0001
Concreteness	-0.49	-0.36	-4.64*	<0.0001	-0.48	-0.32	-6.57*	<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	0.01	0.9913	-0.43	-0.36	-2.21	0.0273
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 and clinicians gave consent to the use of their data in a de-identified, aggregate format as part of the user agreement before they began using the platform and can opt out at any time by informing their therapist or by contacting support. Study procedures were approved as exempt by the our institution's Institutional Review Board (IRB).

Transcripts were de-identified algorithmically via a HIPAA-compliant interface by anonymizing all proper nouns, places, persons, and other nominal features of language. All information related to forms of contact are also removed, including emails, phone numbers, addresses, though these were infrequently found in the interaction between therapists and patients.

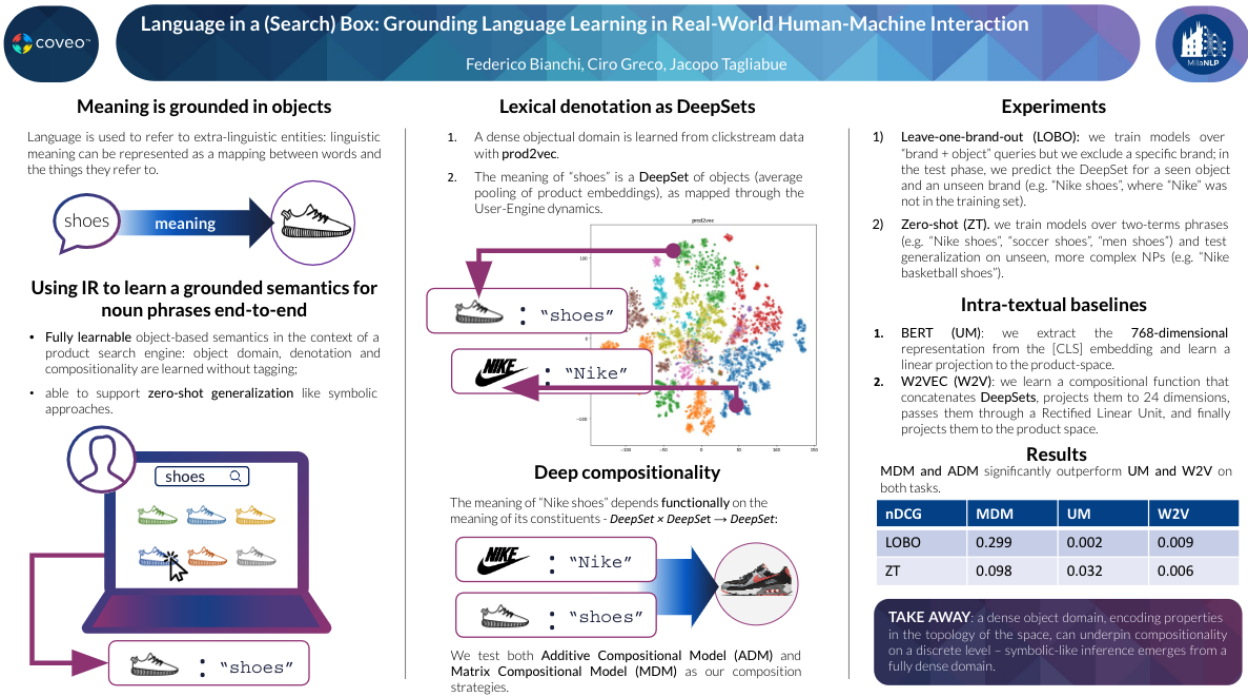
Ethical Considerations

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 Dartmouth. All patients and clinicians 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 informing their therapist or contacting support. We emphasize that the findings in our paper are specific to this dataset and we make no claims about their generalizability to other contexts. Our study was a non-clinical investigation of the complexity loss paradox in psychology, as opposed to a psychiatric study designed for clinical or practical applications. Finally, the data (text messages) were within in English and therefore we do not claim that our findings generalize to other languages. For these reasons, we advise caution when working in this domain 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.

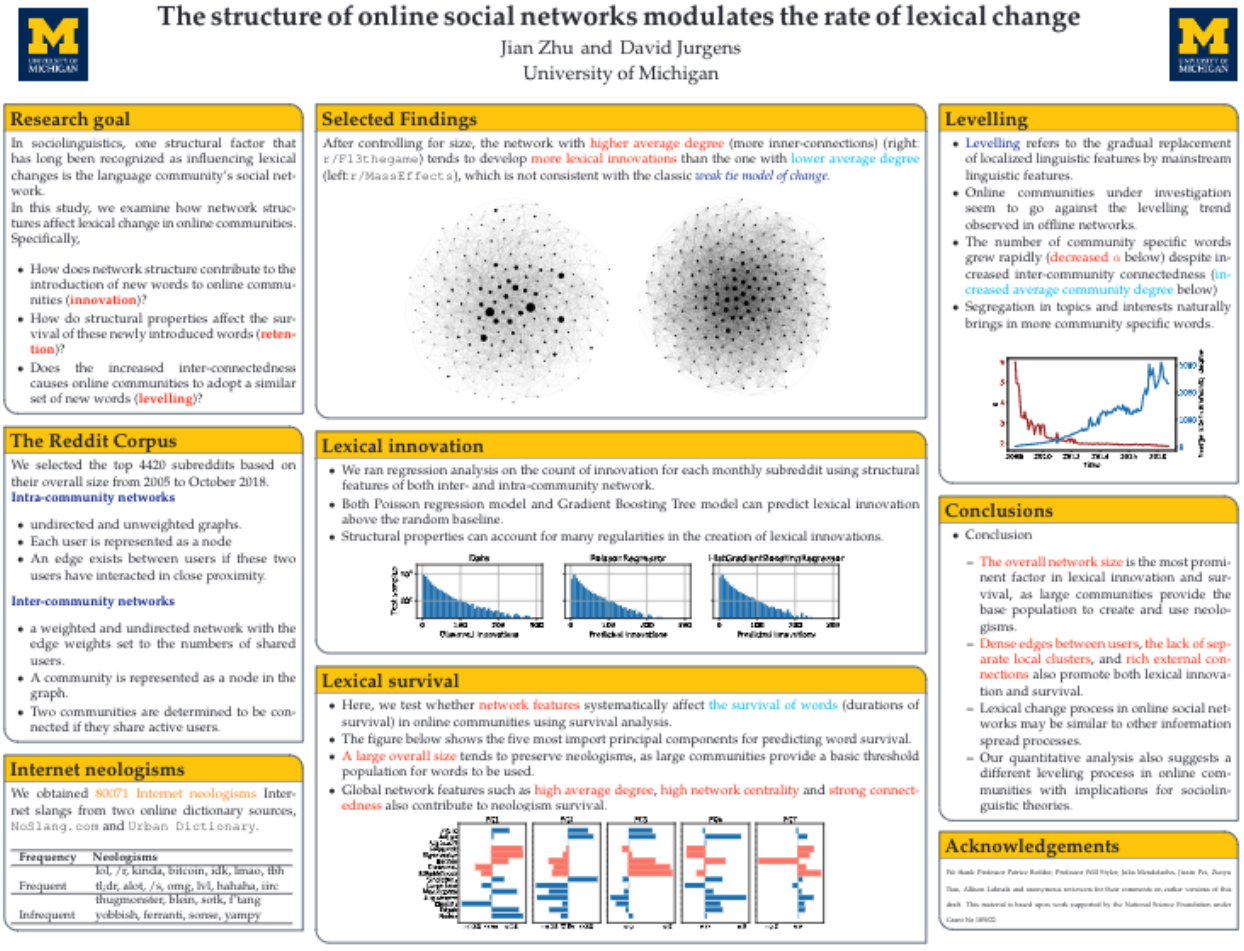
Language in a (search) box: grounding language learning in real-world human-machine interaction



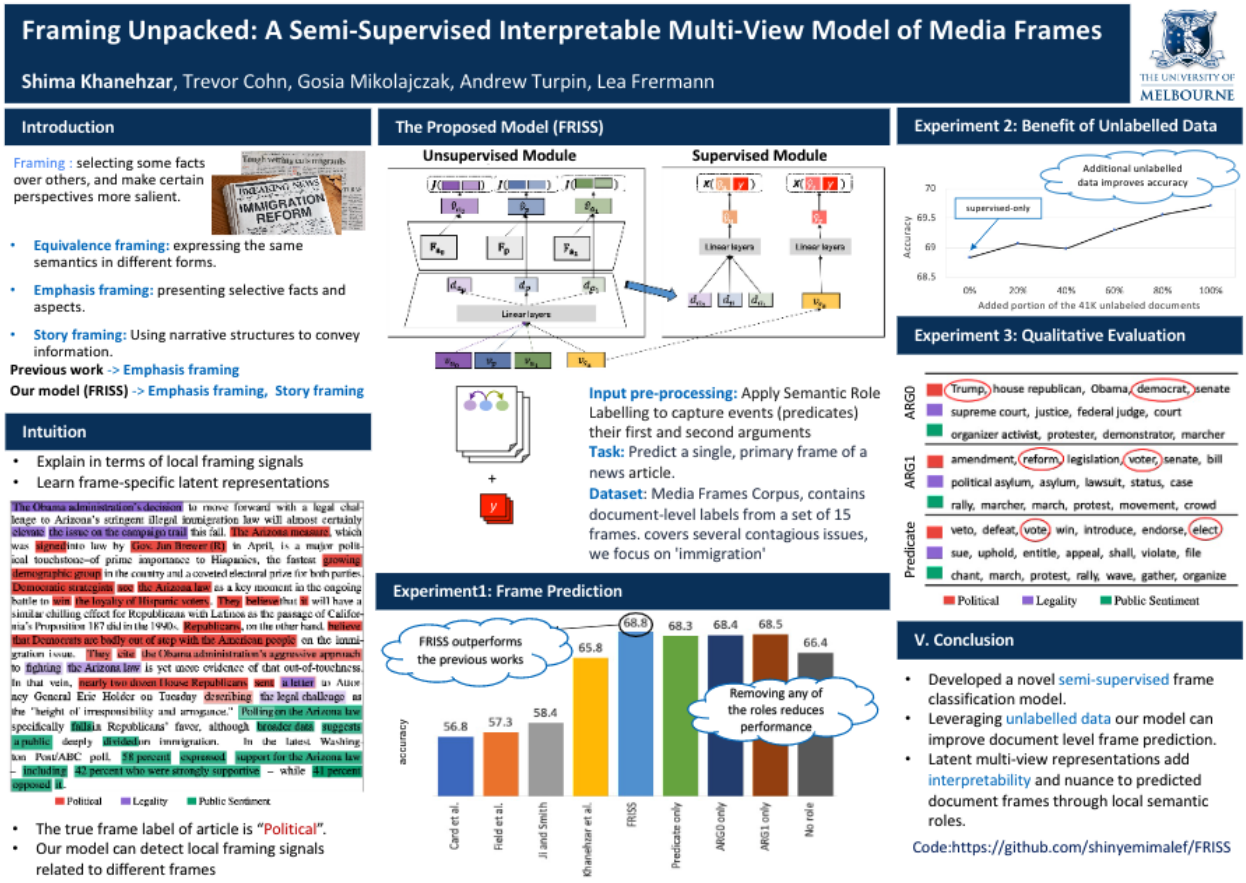
Papers in Computational Social Science

Session 6E

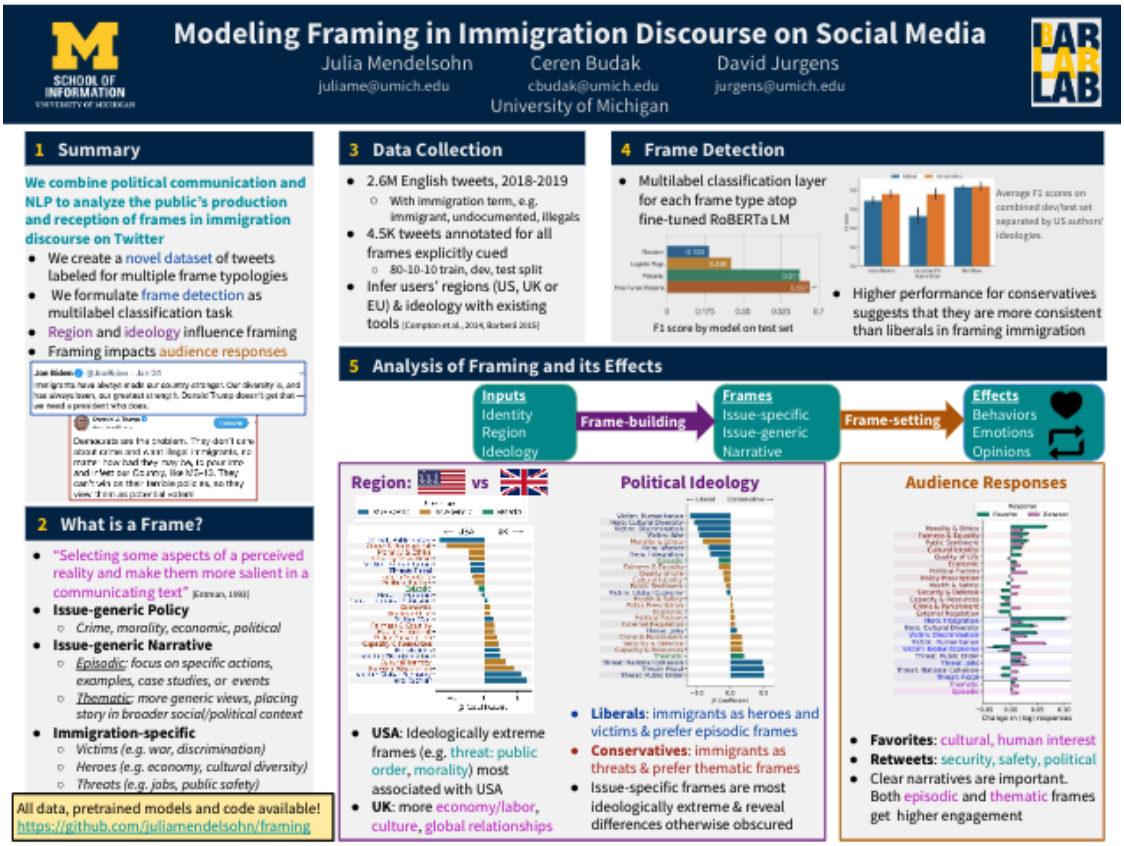
The structure of online social networks modulates the rate of lexical change



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

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

WikiTalkEdit: a dataset for modeling editors' behaviors on Wikipedia

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

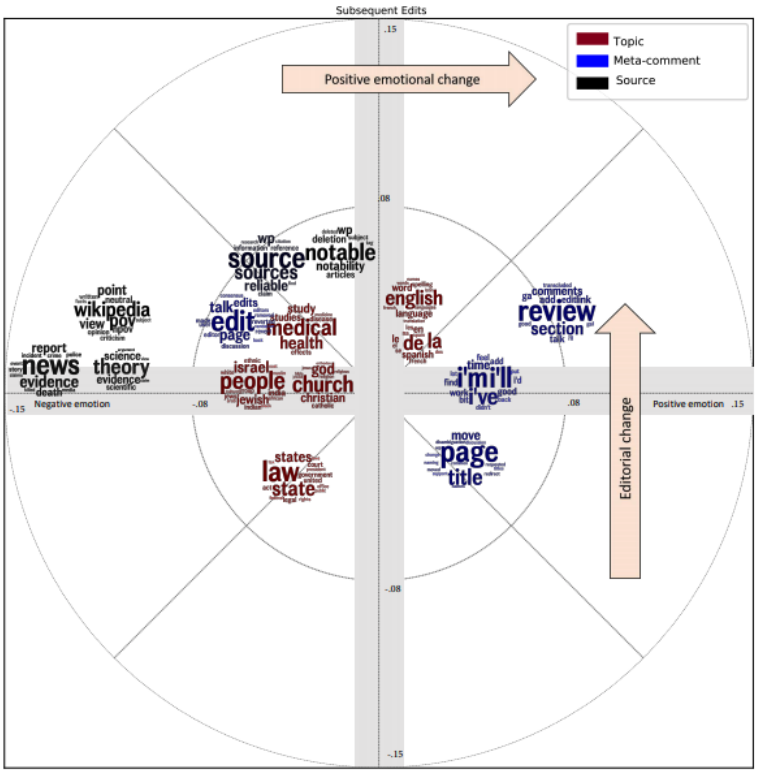
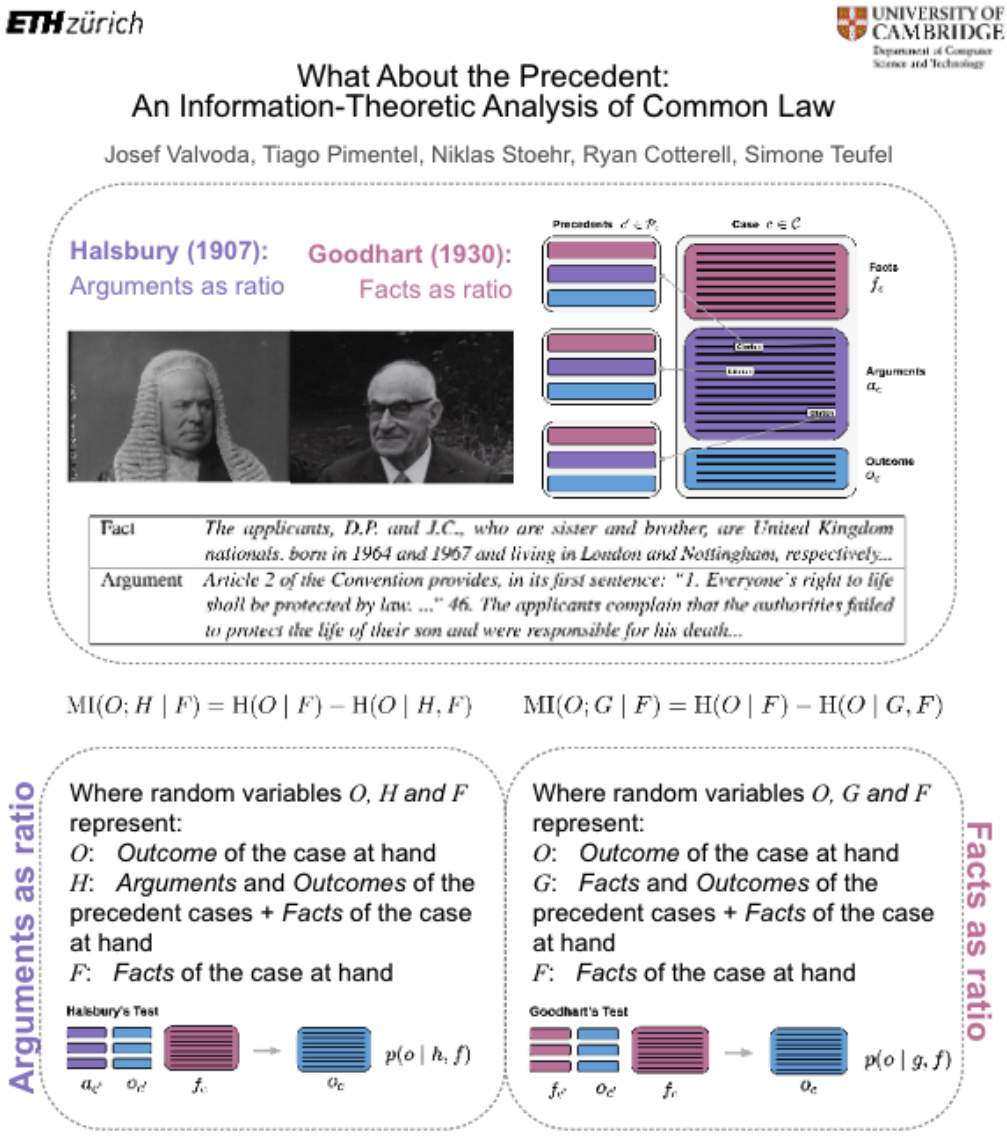


Figure 2: LDA topics correlated with emotional and editorial change. Topics are colored according to their theme; word size is proportional to word weight in the topic.

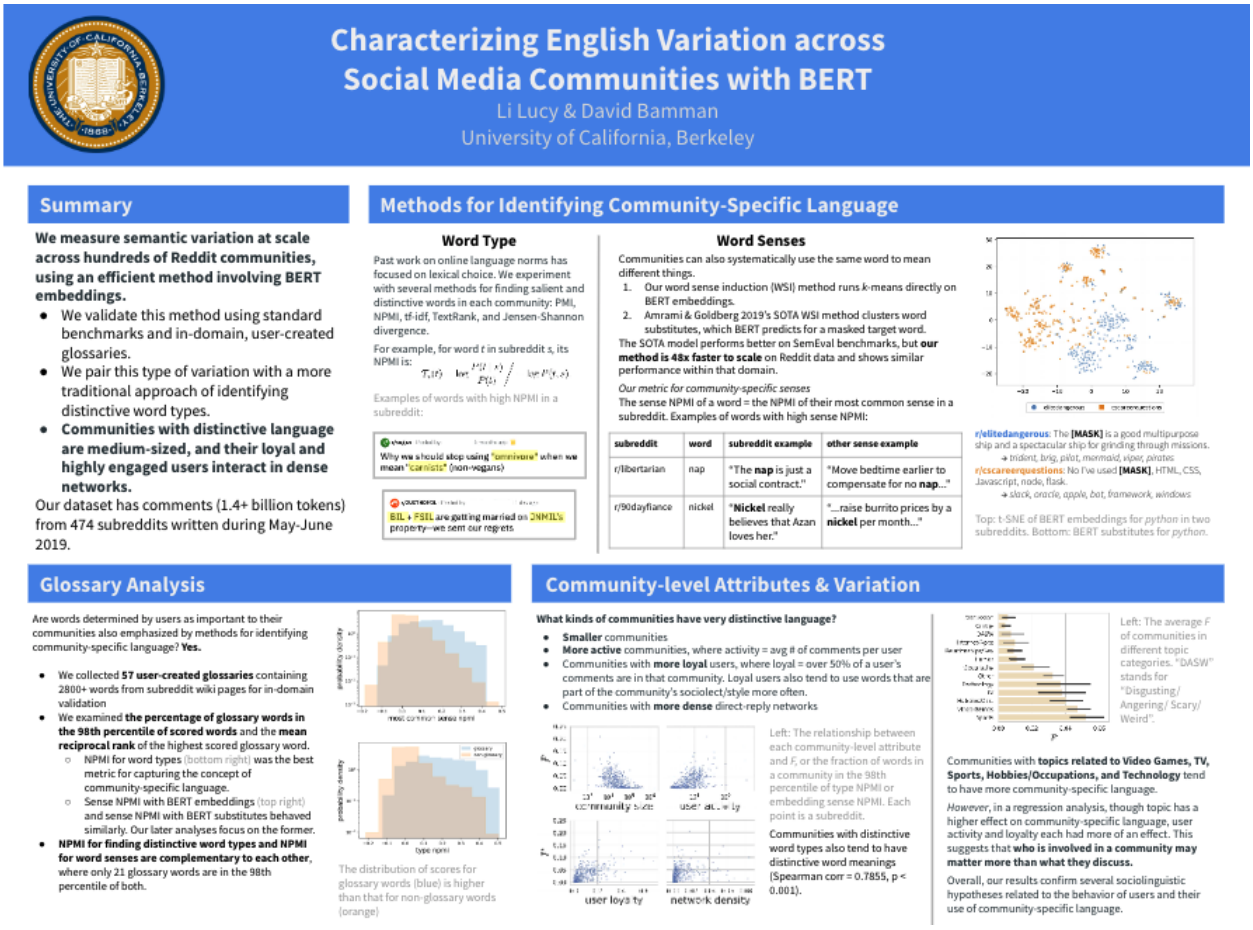
- More descriptions about dataset in paper.

Session 7A

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

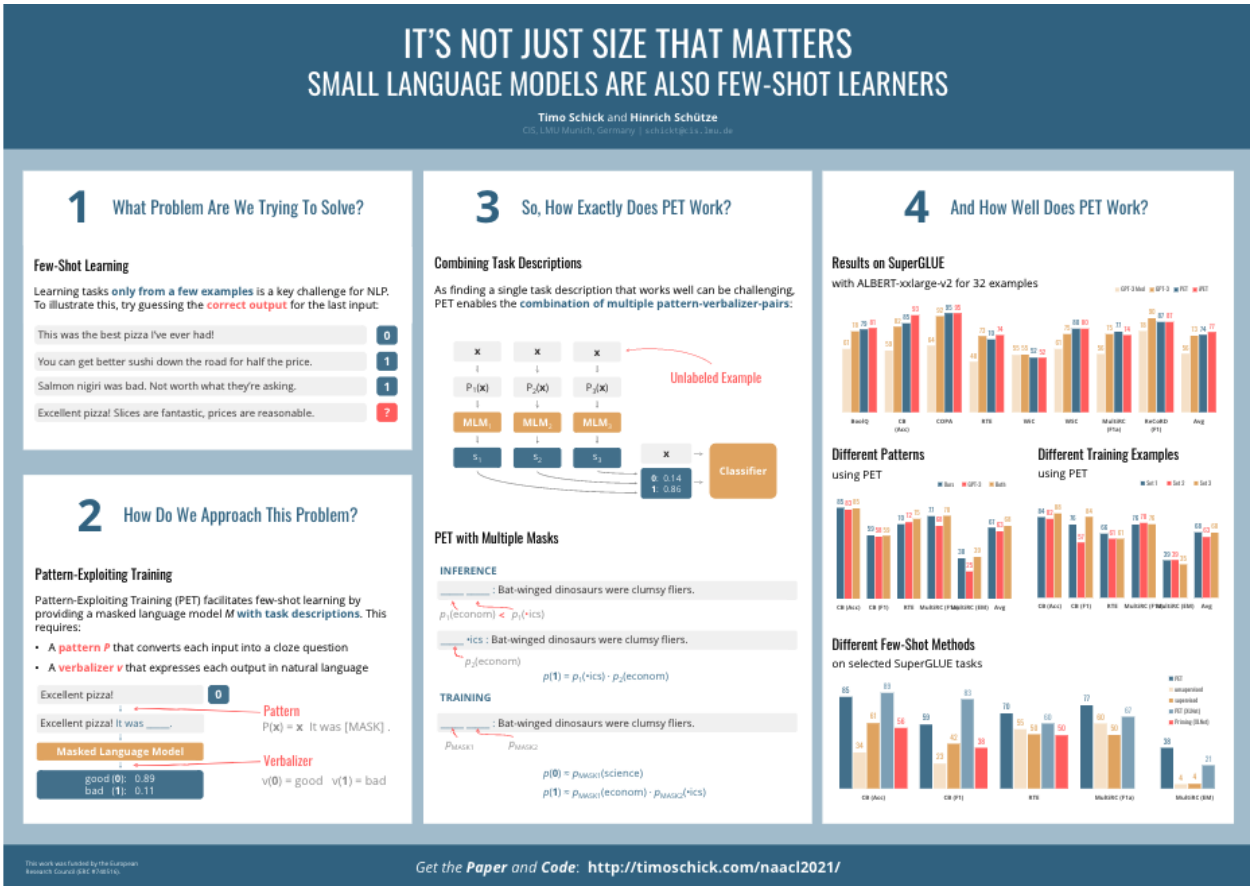


Characterizing English variation across social media communities with BERT



Session 7B (Green NLP)

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



Static embeddings as efficient knowledge bases?

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

Comparison

Model	Vocab. Size	p1 LAMA	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

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

Typed Querying

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

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

Advantages:

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

Evaluation

We compute precision at one (p1) for each relation, i.e., $1/|T| \sum_{t \in T} \mathbb{1}\{\hat{t}_{object} = t_{object}\}$ where T is the set of all triples and $\hat{t}_{object}$ is the object predicted.
Dataset: TREx with 41 relations; 10 languages.

Multilingual Results

Model	Vocab. Size	p1 AR	DE	ES	FI	HE	JA	KO	TH	TR
Oracle		21.9	22.3	21.6	21.3	22.9	21.3	21.7	23.7	23.5
mBERT	110k	17.2	31.5	33.6	20.6	17.5	15.1	18.9	13.5	33.8
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

Resource Consumption

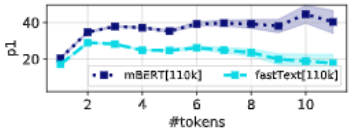
- Static embeddings much **cheaper** to compute
- **0.3%** of carbon emissions of BERT
- Only **CPU** required

Prediction Diversity

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

⇒ fastText predictions are **more diverse**

Contextualization



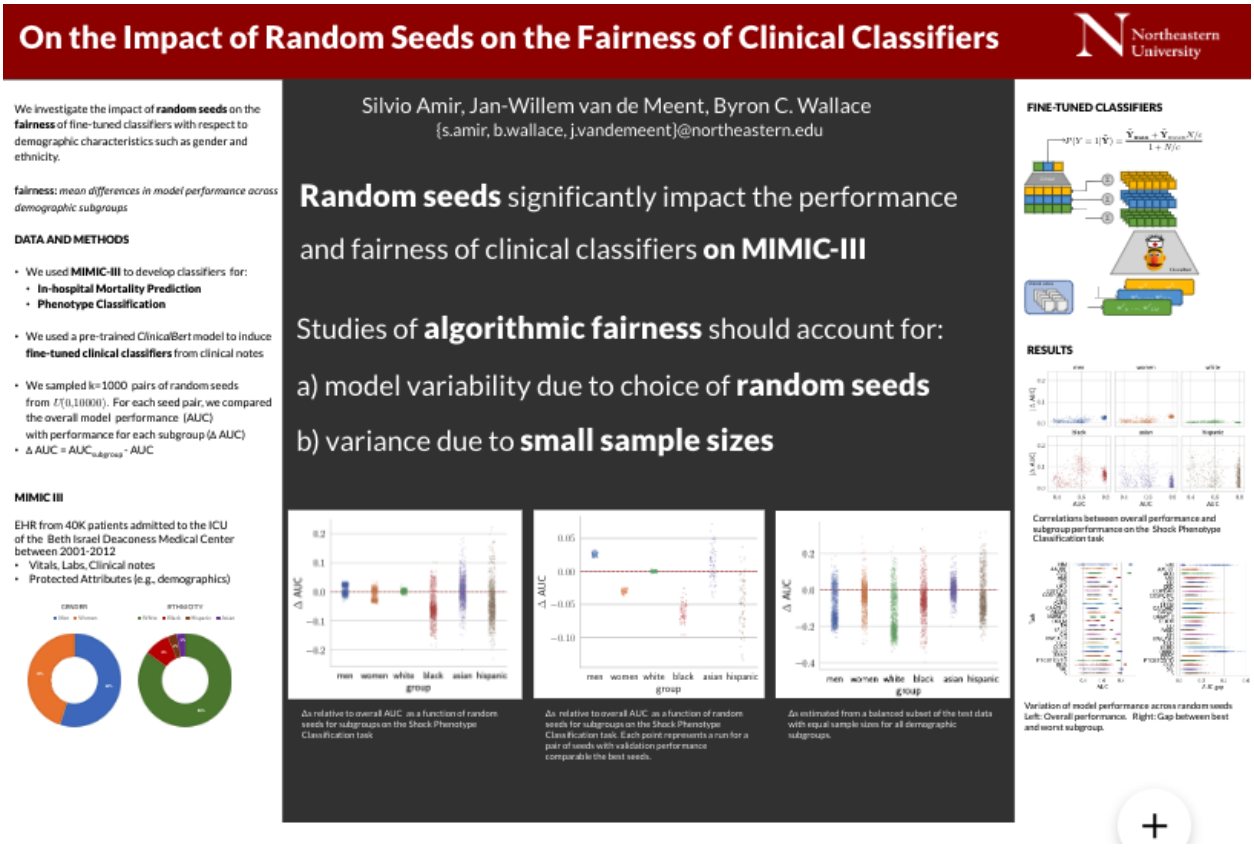
⇒ mean pooling worse for multitoken subjects
⇒ **trade-off**: learning good composition function with small vocabulary vs. large number of atomic representations

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

Presented at NAACL 2021, Online. * Equal Contribution

Session 11A (ethics)

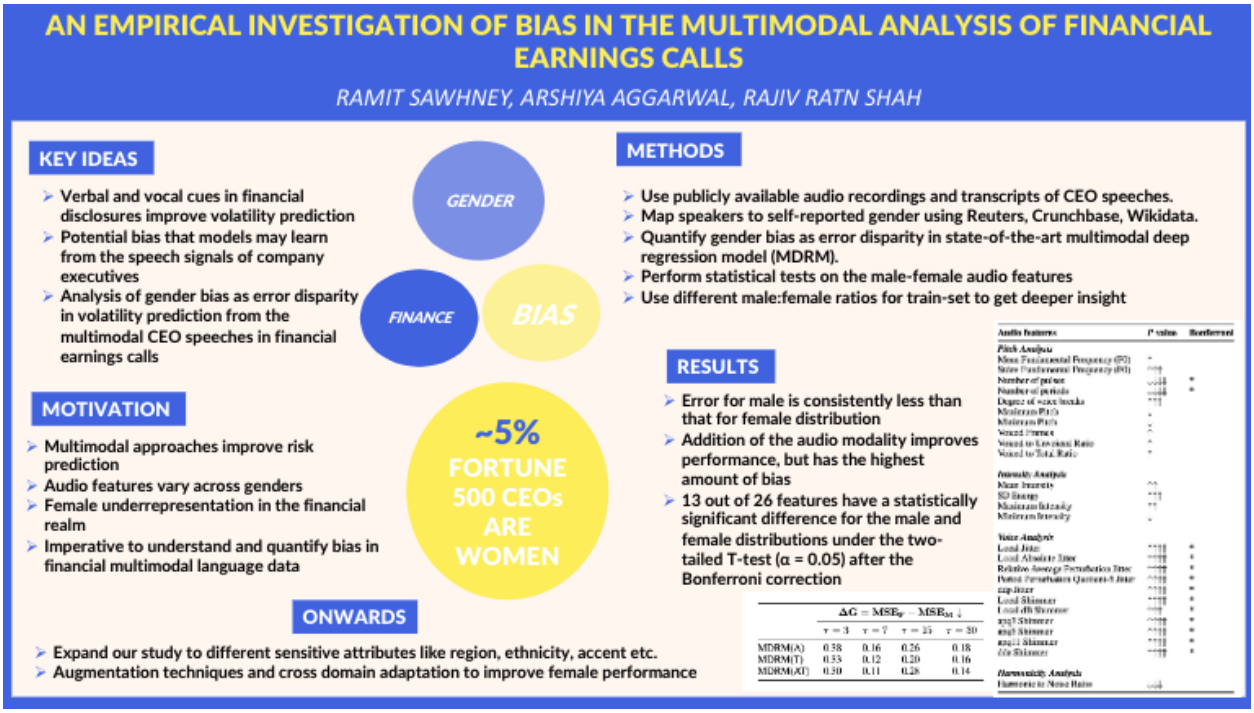
On the impact of random seeds on the fairness of clinical classifiers



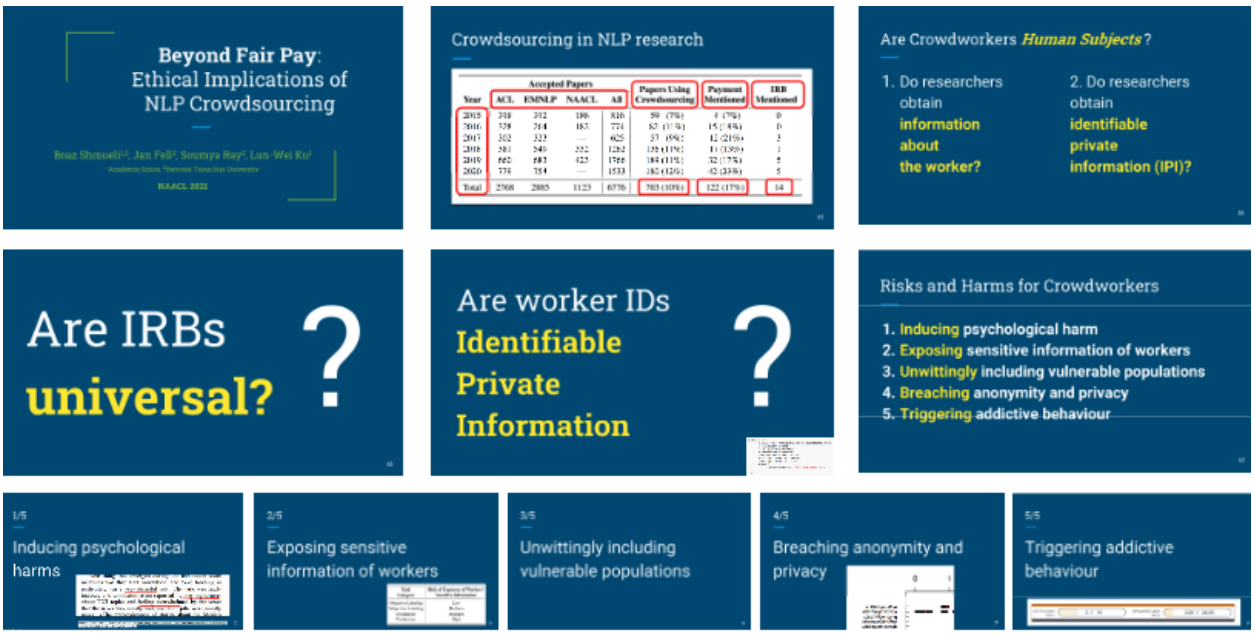
Dynamically disentangling social bias from task-oriented representations with adversarial attack

- 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).

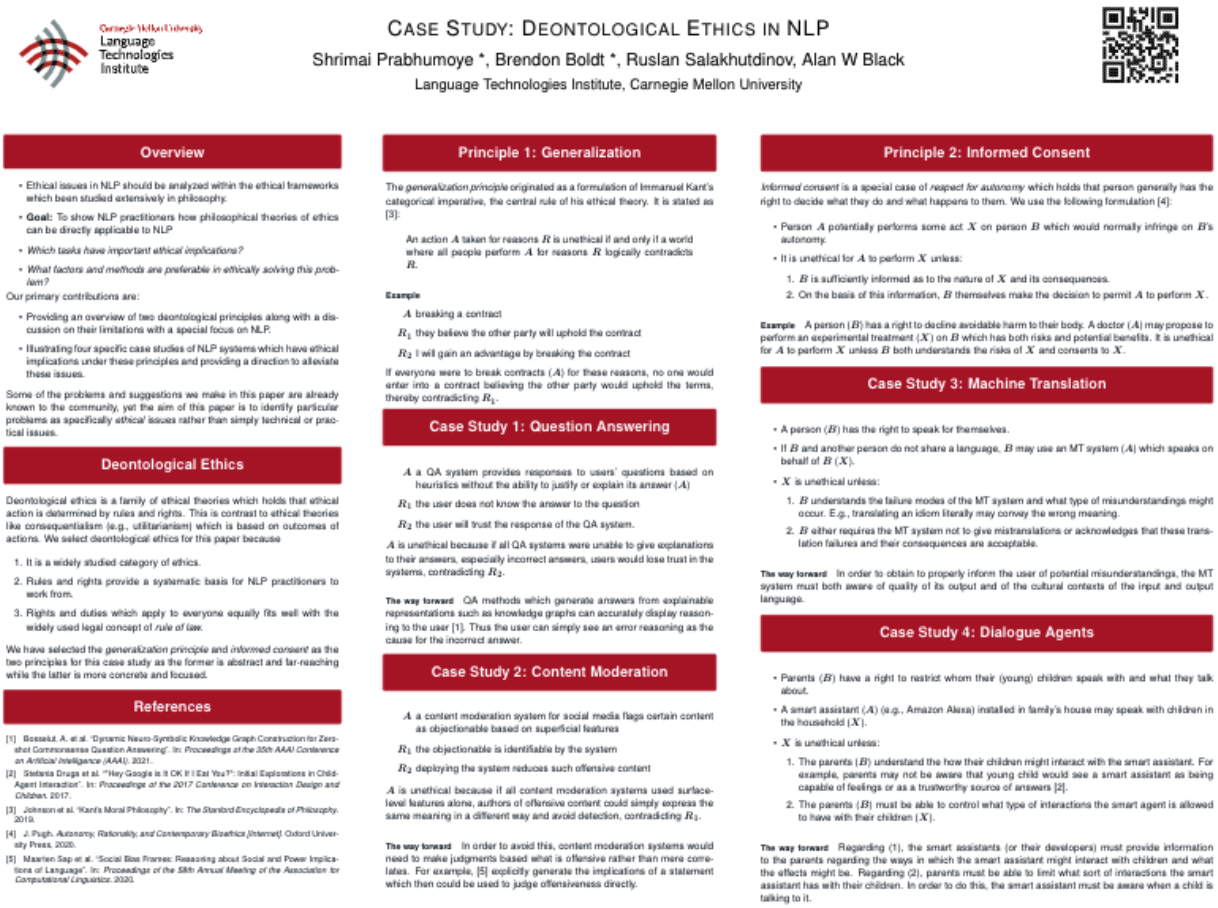
An empirical investigation of bias in the multimodal analysis of financial earnings calls



Beyond fair pay: ethical implications of NLP crowdsourcing

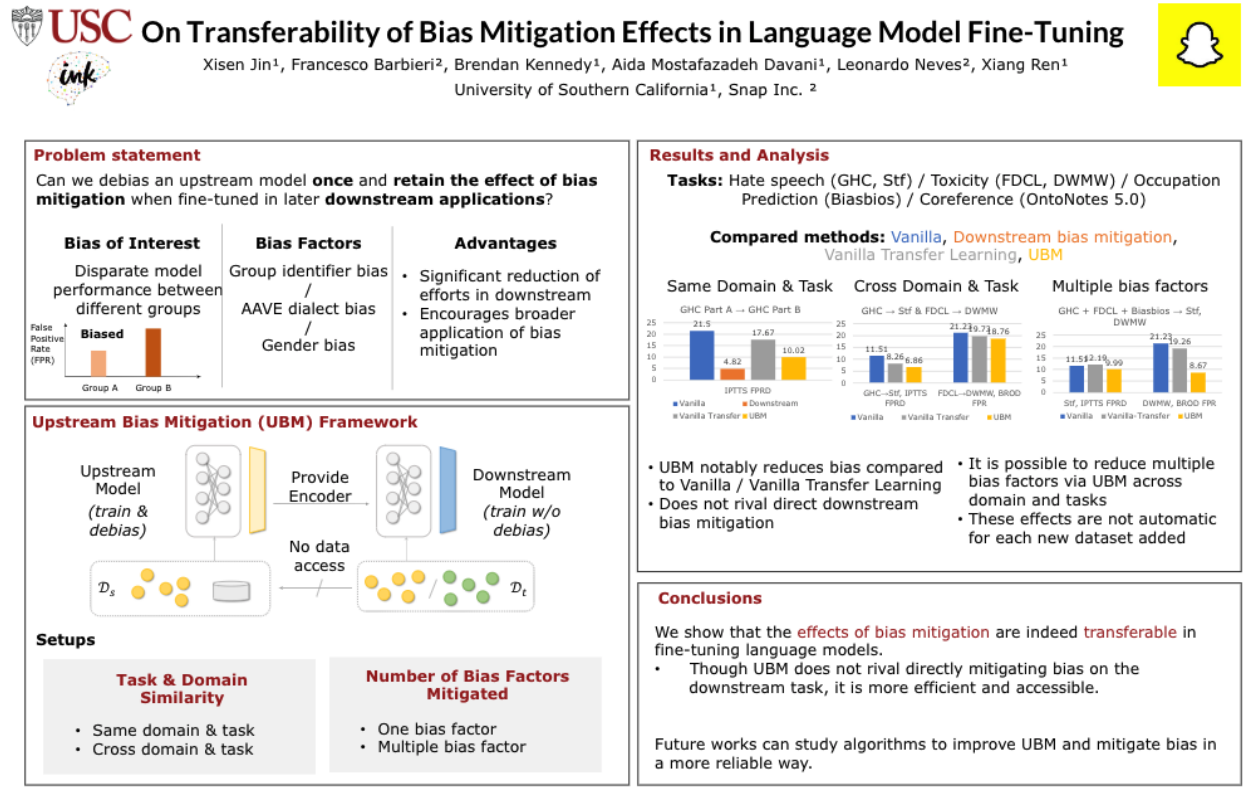


Case study: deontological ethics in NLP

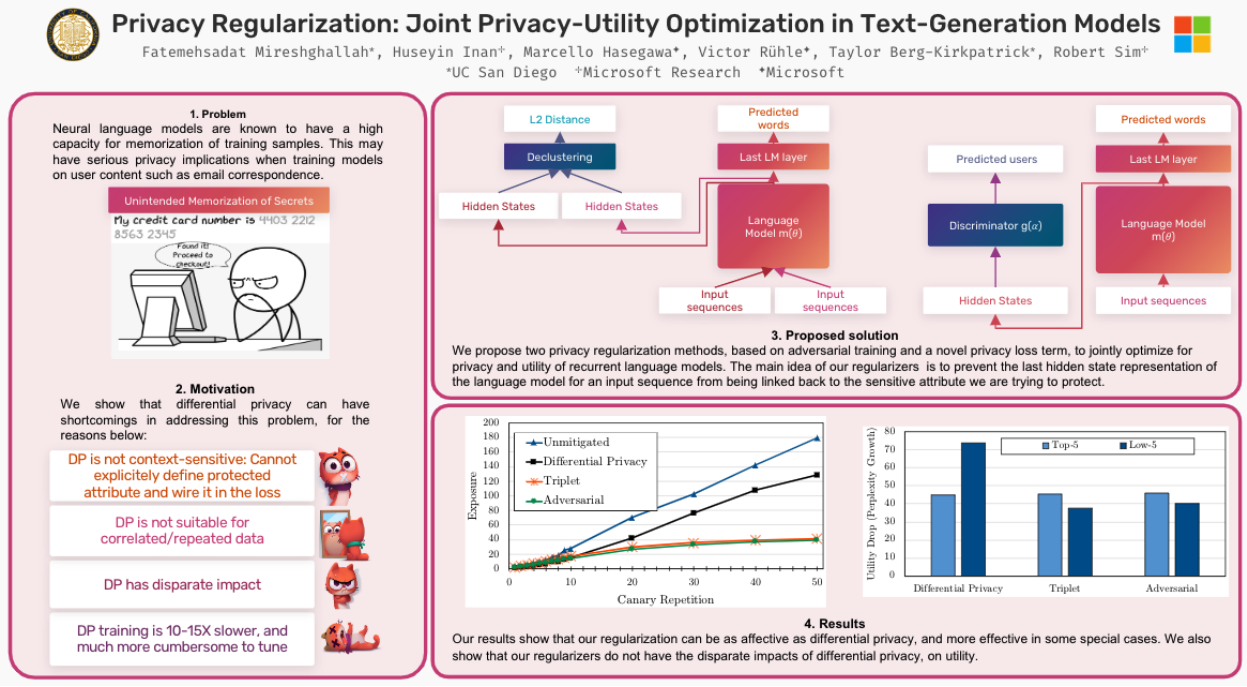


- Goal: leverage a large body of work on ethics. See how we can apply them to NLP.
- Deontological framework for NLP
  - Generalization principle (categorical imperative: An action A is ethical *iff* a world where all people performing A is conceivable)
  - Respect for Autonomy
- Reasonable, clear ethical 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



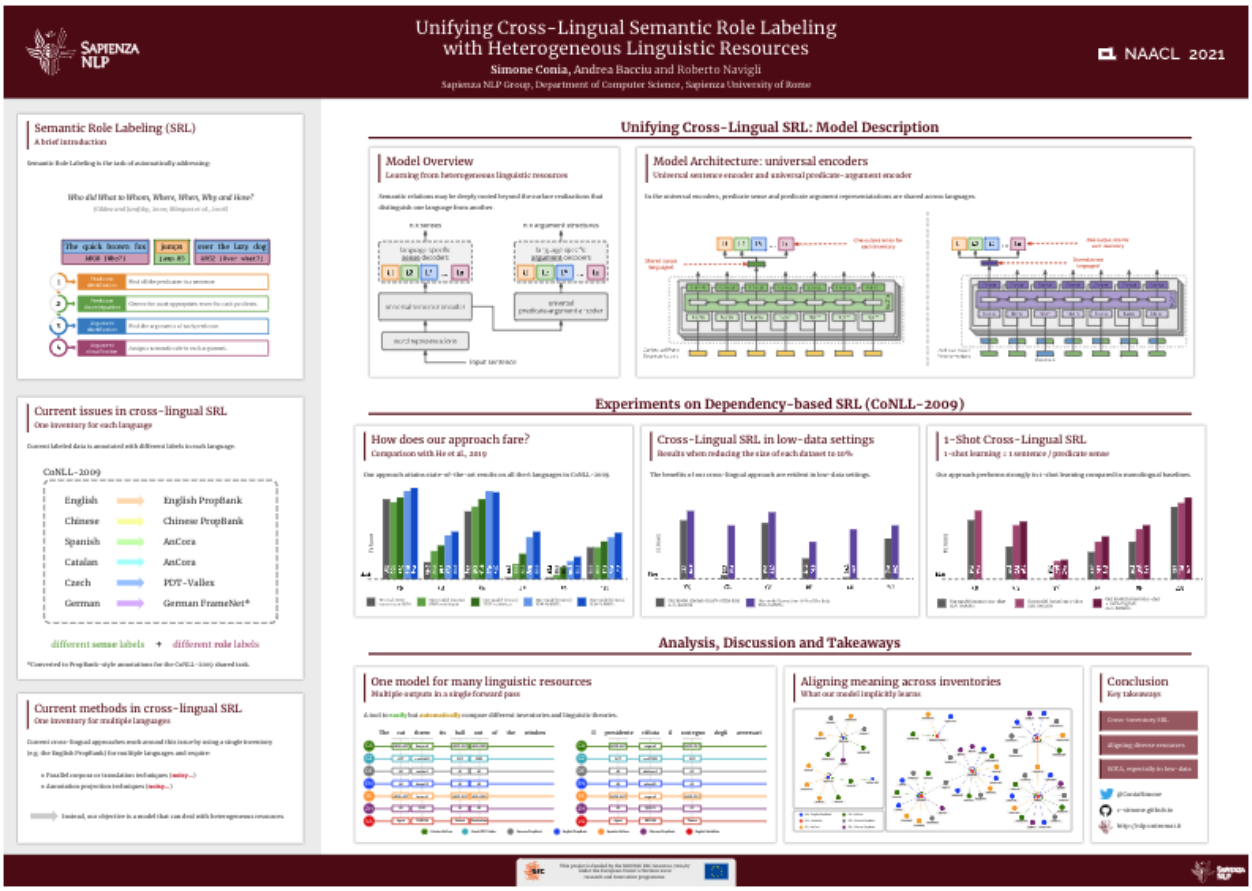
Privacy regularization: joint privacy-utility optimization in LMs



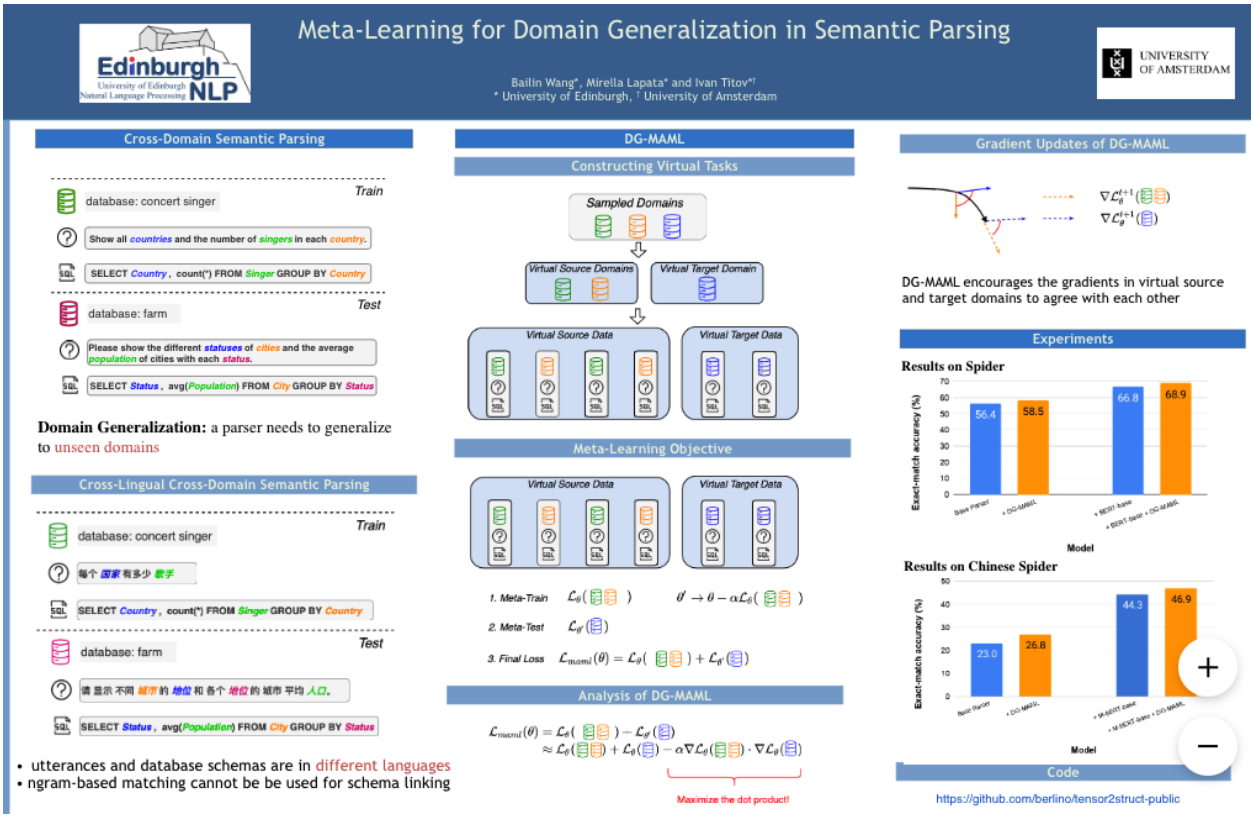
Papers in semantics

Session 1E (sentence-level, textual inference)

Unifying cross-lingual SRL with heterogeneous linguistic resources

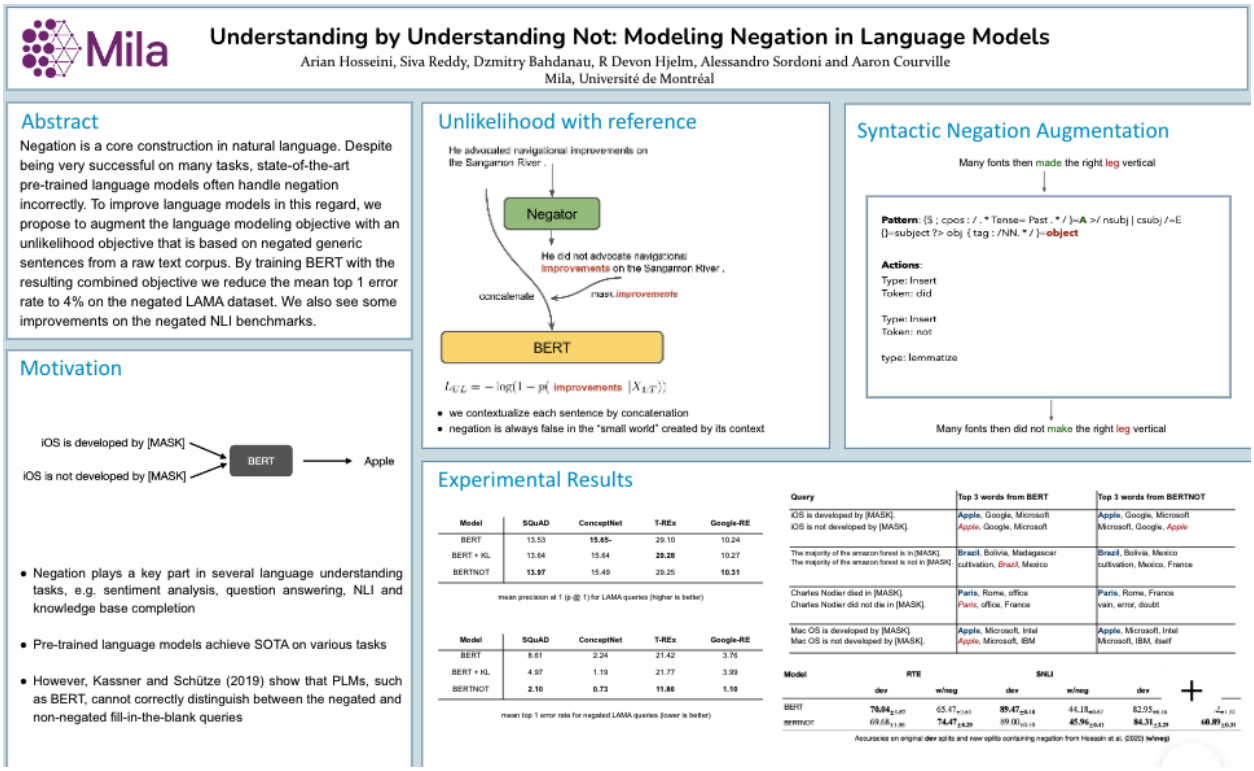


Meta-learning for domain generalization in semantic parsing

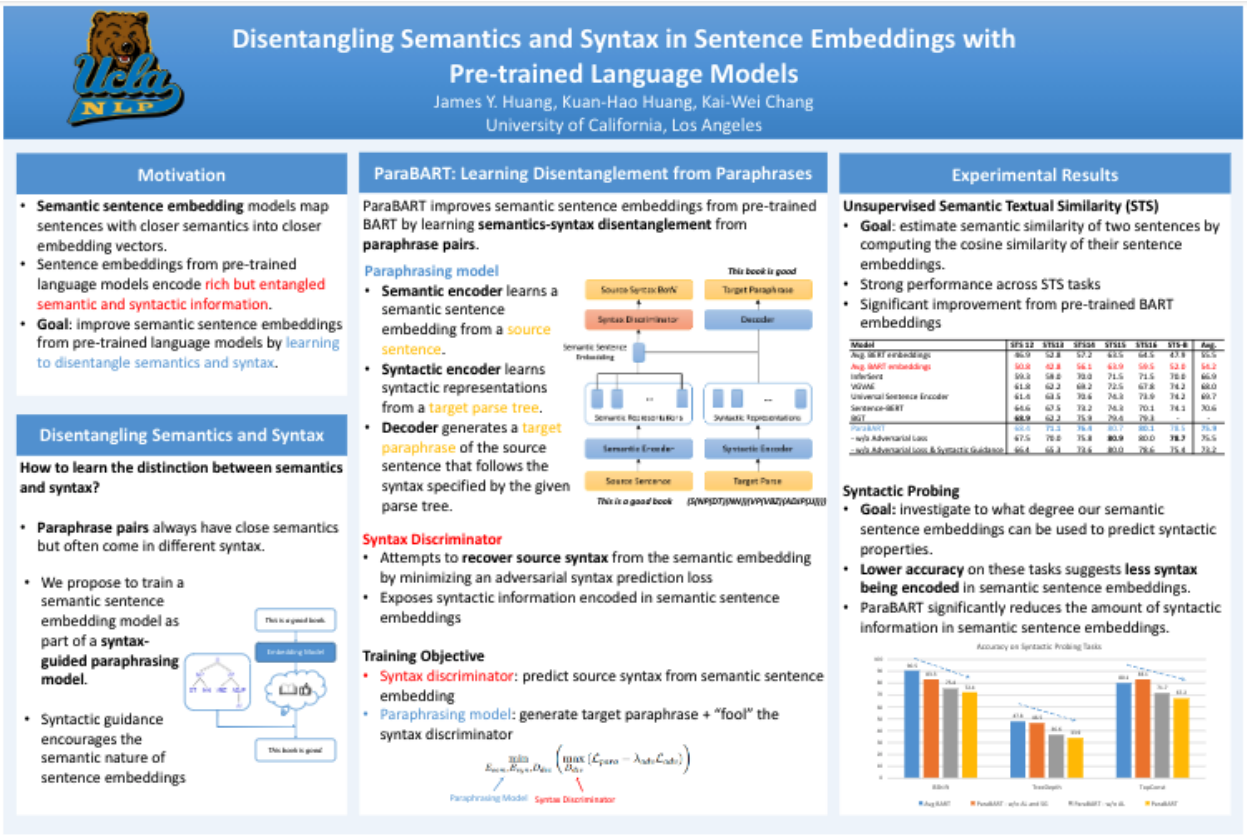


Session 4C (sentence-level, textual inference)

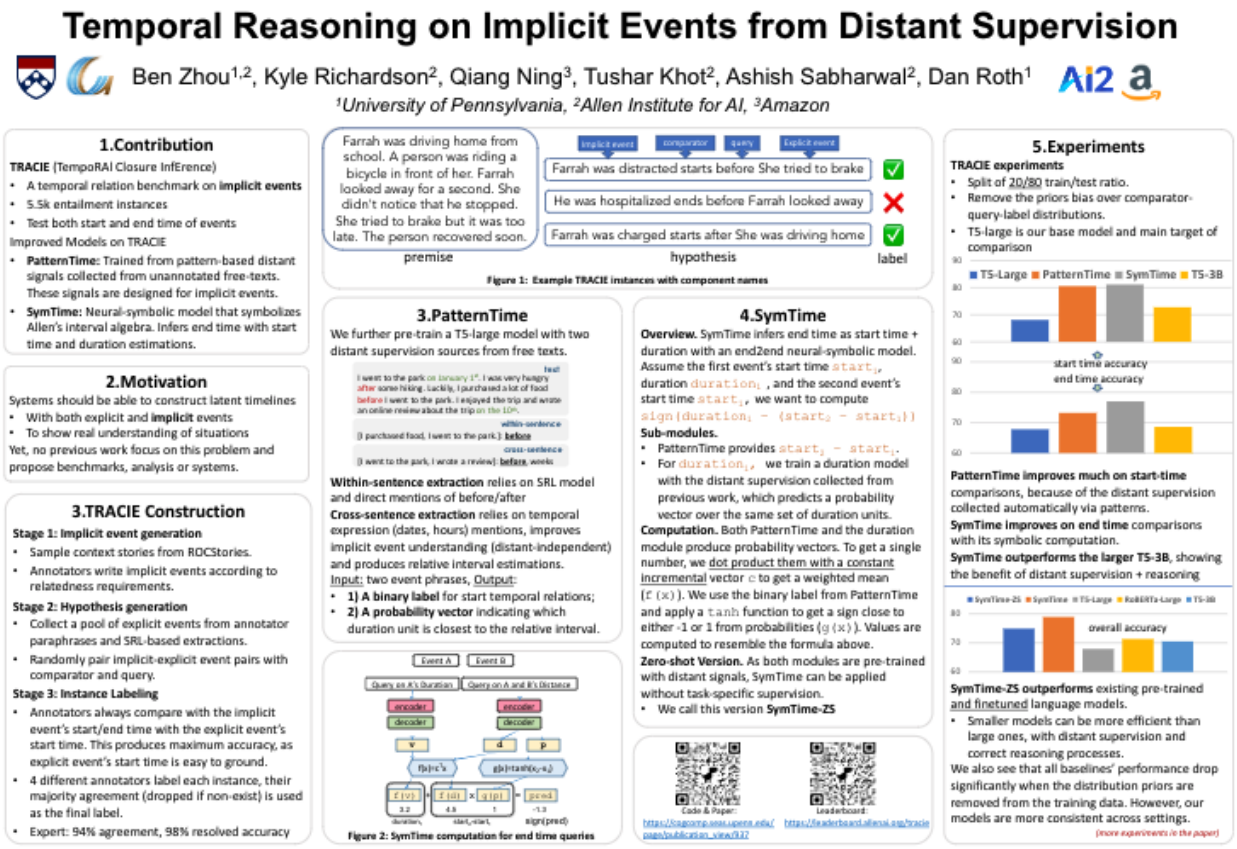
Understanding by understanding not: modeling negation in LMs



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

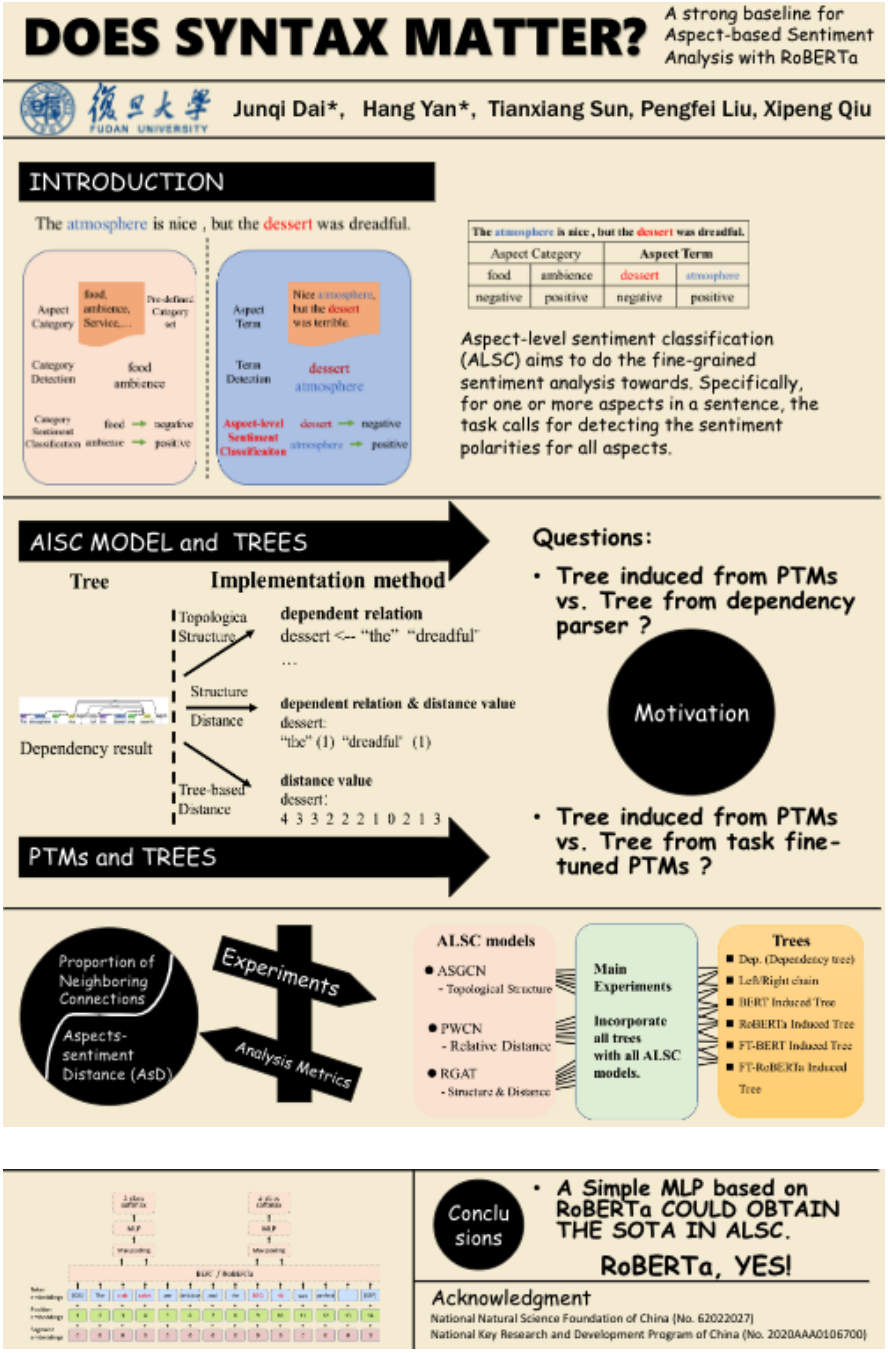


Temporal reasoning on implicit events from distant supervision

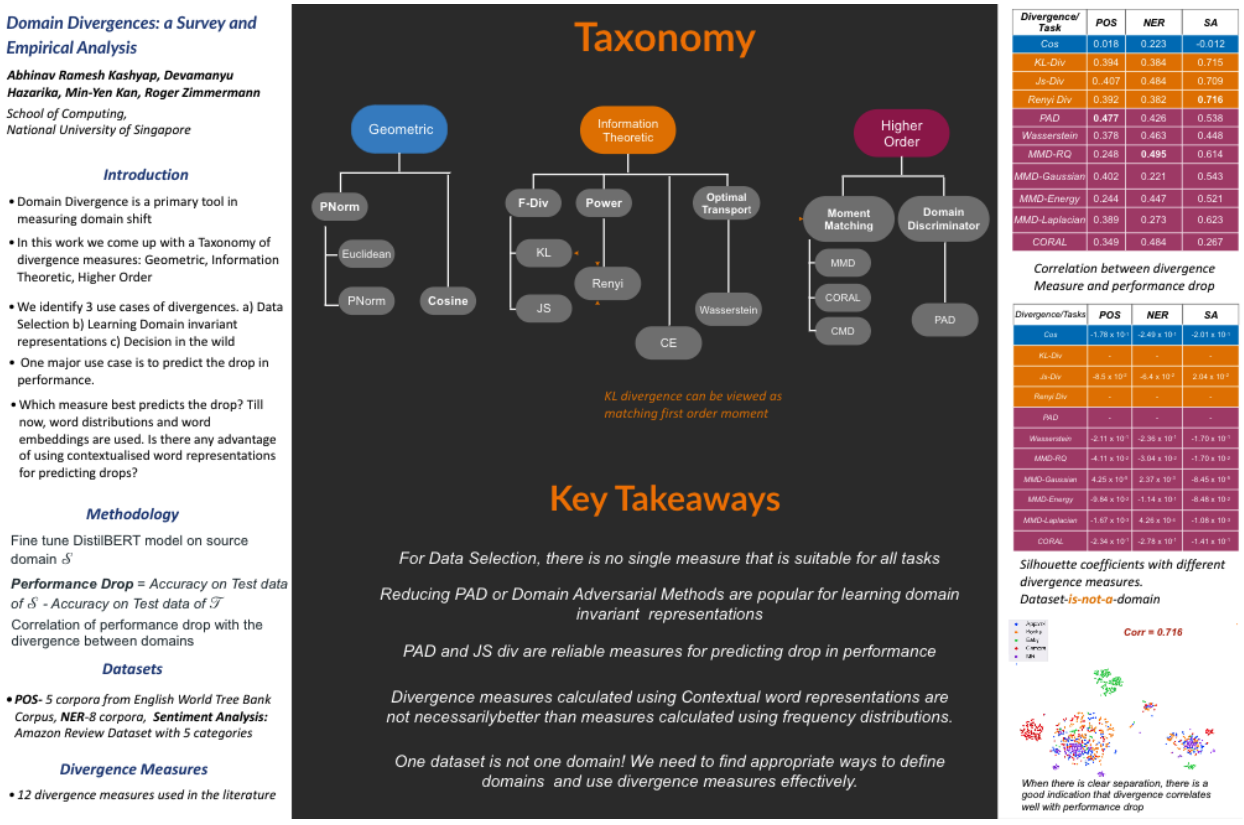


Session 5E (stylistic analysis)

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

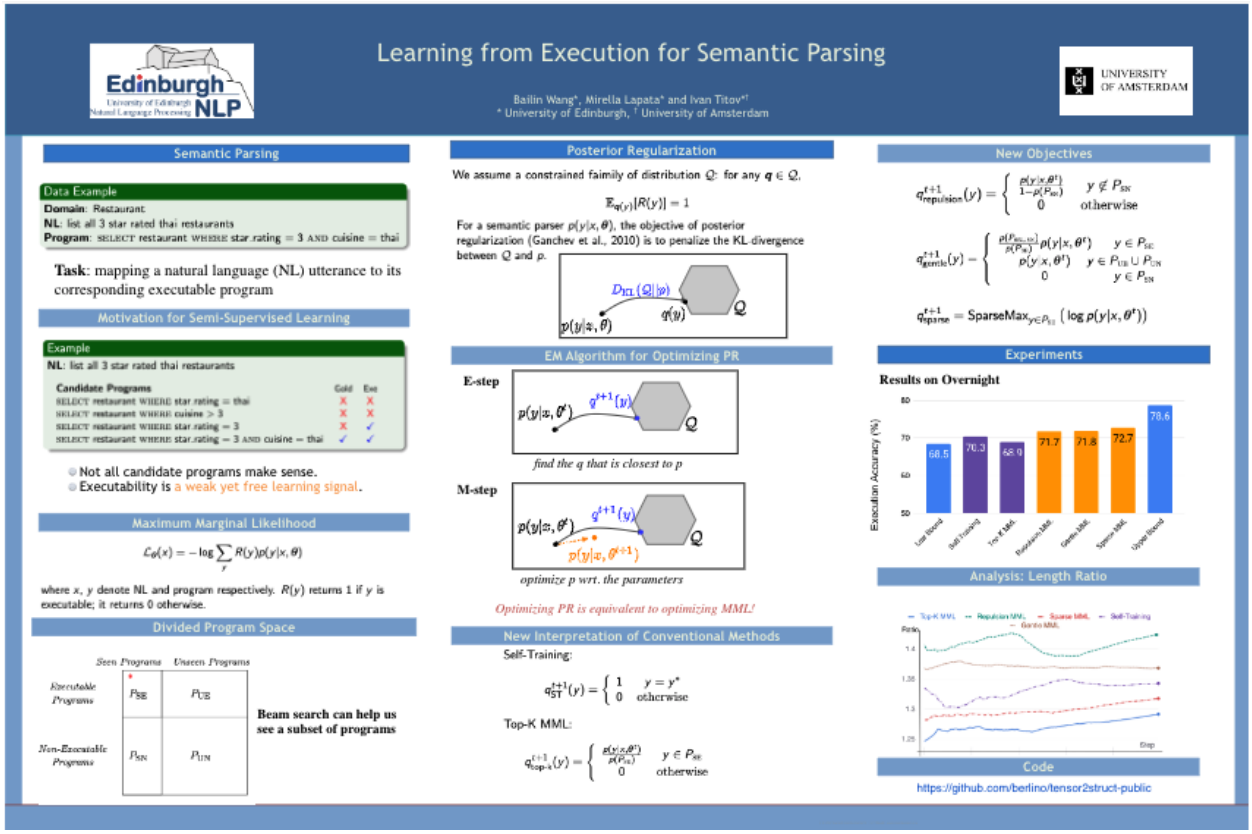


Domain divergences: a survey and empirical analysis

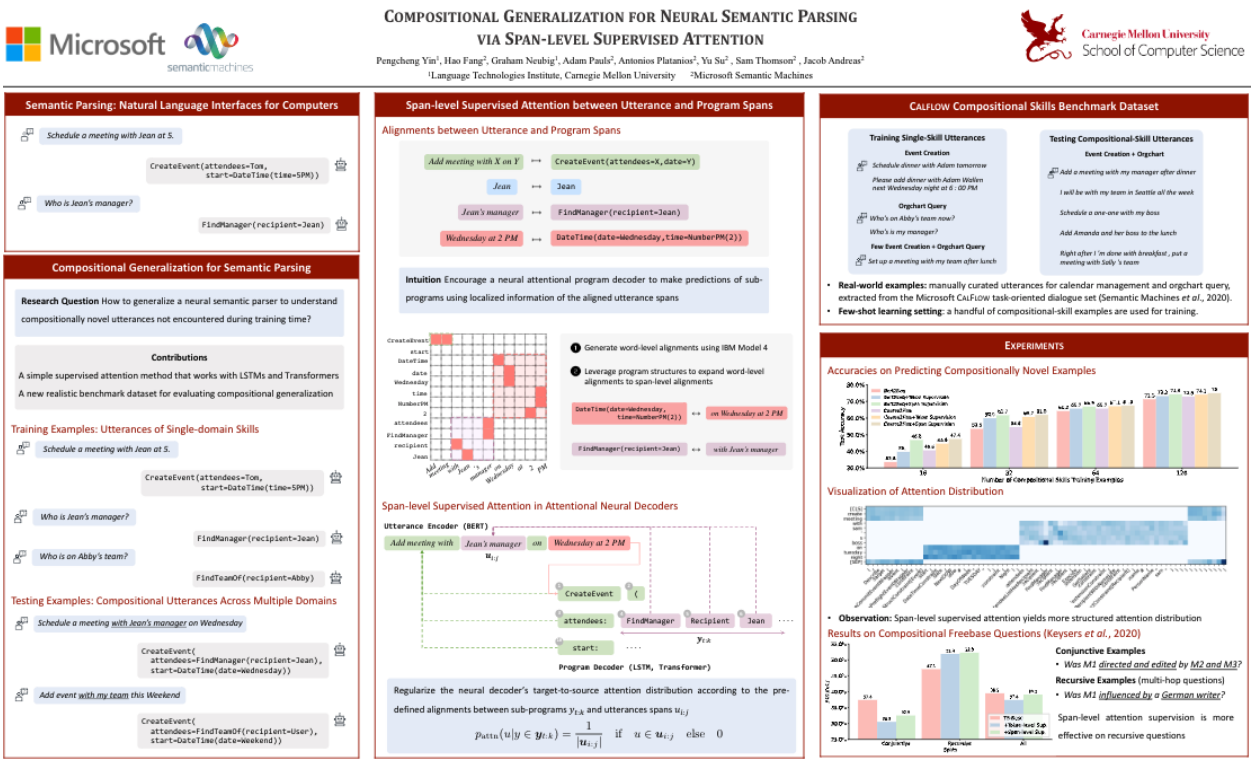


Session 8C (sentence-level, textual inference)

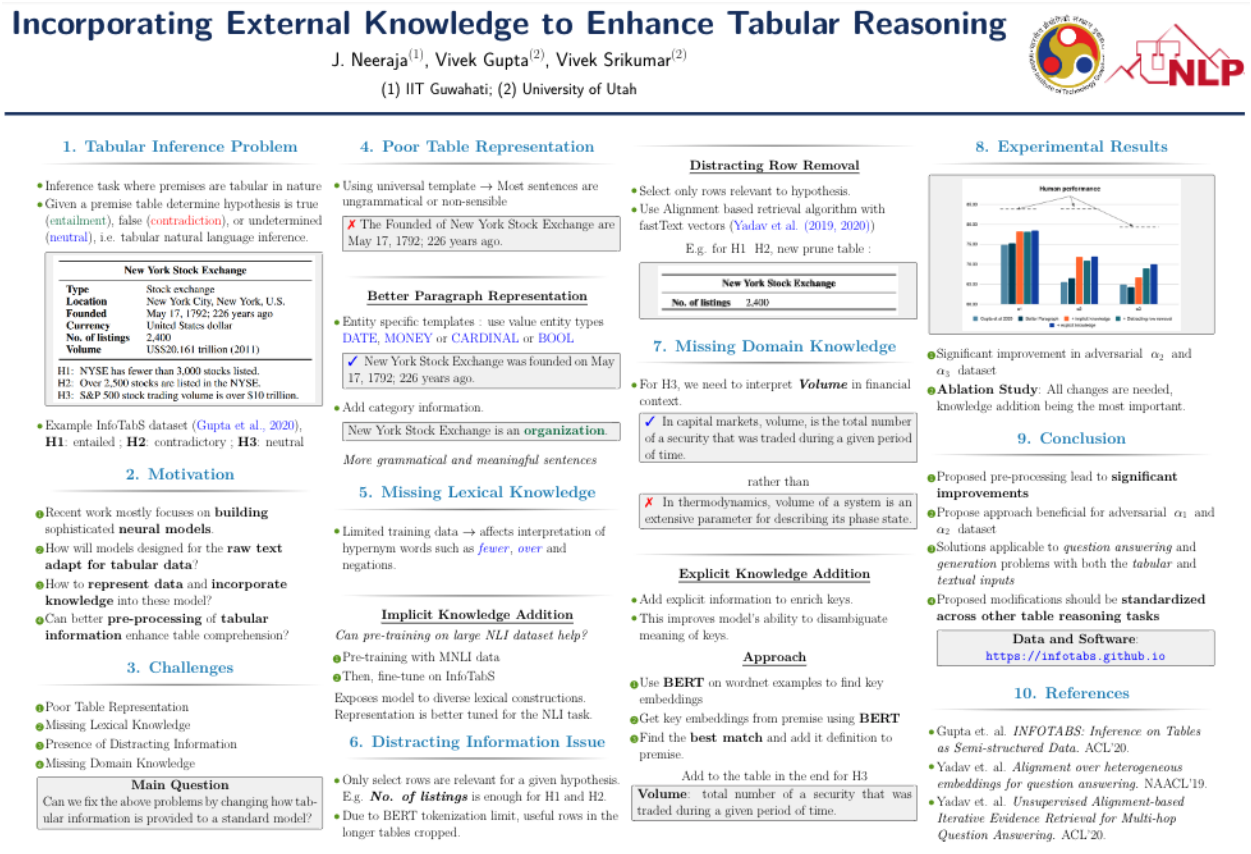
Learning from executions for semantic parsing



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

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

BROWN

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

Inputs: NLP dataset D with full vocabulary V
for each word w in V **do**
  $\phi_w \leftarrow 0$ (initialise Shapley value estimate)
  for i=1 to S (number of sampled permutations **do**
    $\pi \leftarrow$ Random-Permutation(V)
    $C_1 \leftarrow b(w, \pi)$ (predecessors of w)
    $C_2 \leftarrow C_1 \cup \{w\}$ (predecessors including w)
    $f_\theta^{C_1} \leftarrow$ TrainModel(C_1) (Train on vocabulary C_1)
    $f_\theta^{C_2} \leftarrow$ TrainModel(C_1) (Train on vocabulary C_2)
    $m(w, \pi) \leftarrow q(f_\theta^{C_2}) - q(f_\theta^{C_1})$
    $\phi_w \leftarrow \phi_w + m(w, \pi)$
  end for
  $\phi_w \leftarrow \frac{1}{S}\phi_w$ (average marginal contributions)
end for
Rank words in V based on Shapley estimates π_w
Return top k words in ranking

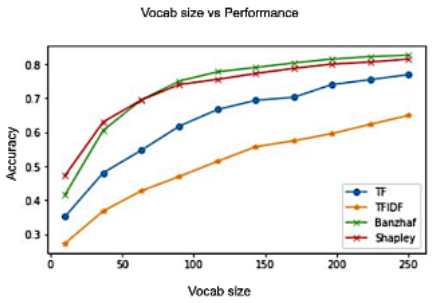


Figure 1:Comparing Shapley, Banzhaf, TF (term frequency) and TF-IDF on AG-news (a document classification task). We see that both game-theoretic algorithms (red and green) outperform the other heuristics for all vocabulary sizes.

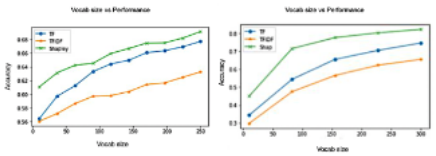


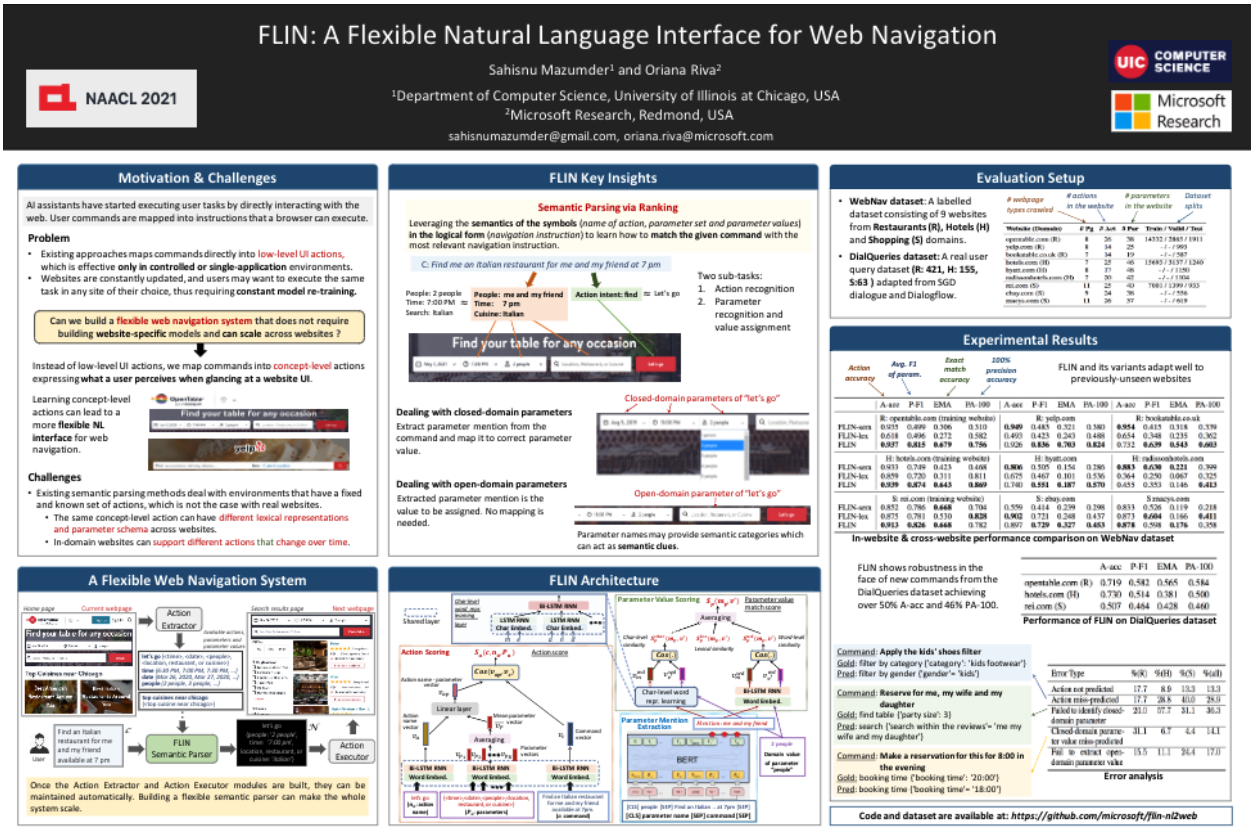
Figure 2:Comparison of Shapley to TF (frequency) and TF-IDF on two additional task structures: single-sentence classification (left) on SST-2 and pairwise-sentence classification (right) on SNLI. We see that Shapley (green) outperforms the baselines in both tasks here, as well as the third task structure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

Evaluation

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

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

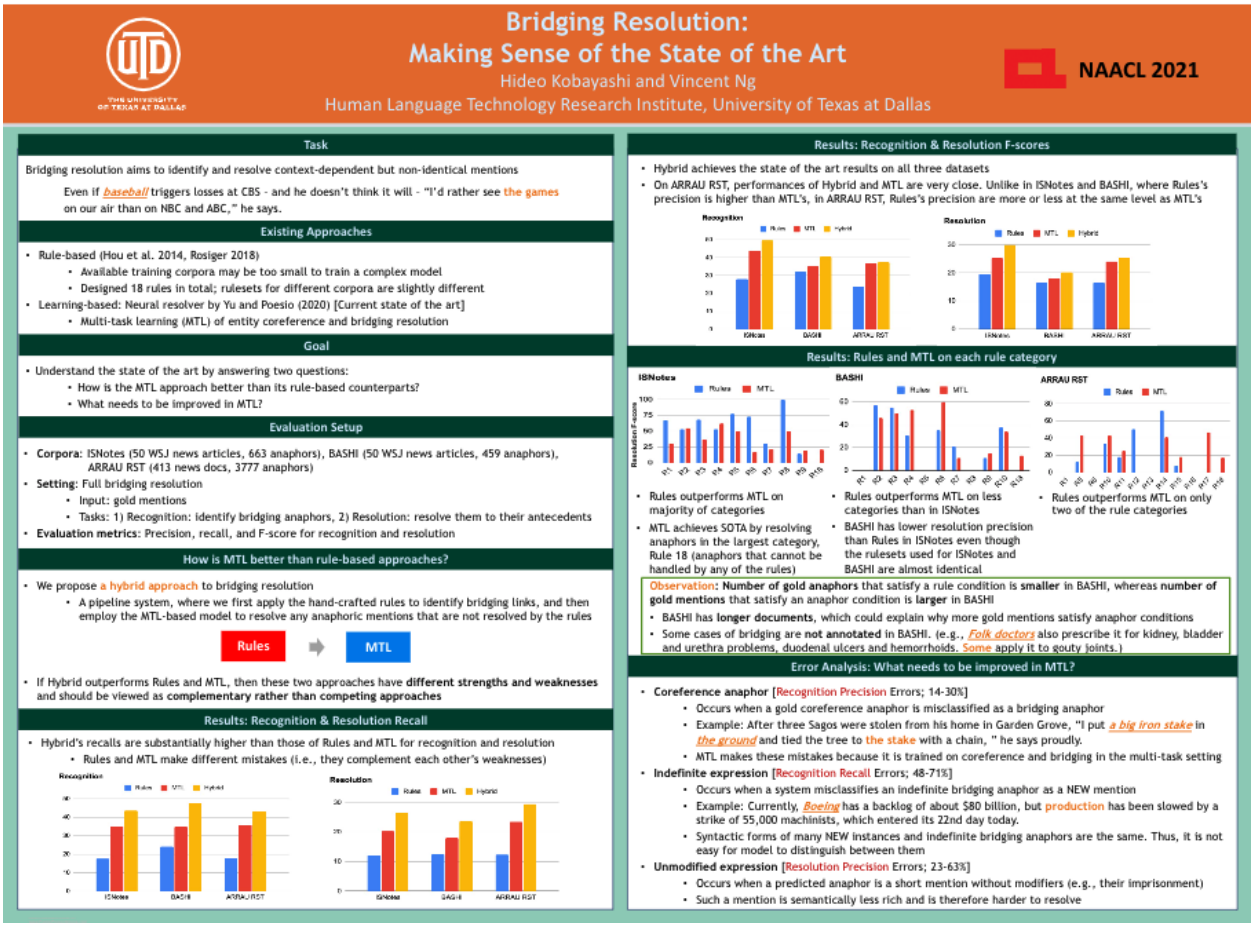
A flexible natural language interface for web navigation



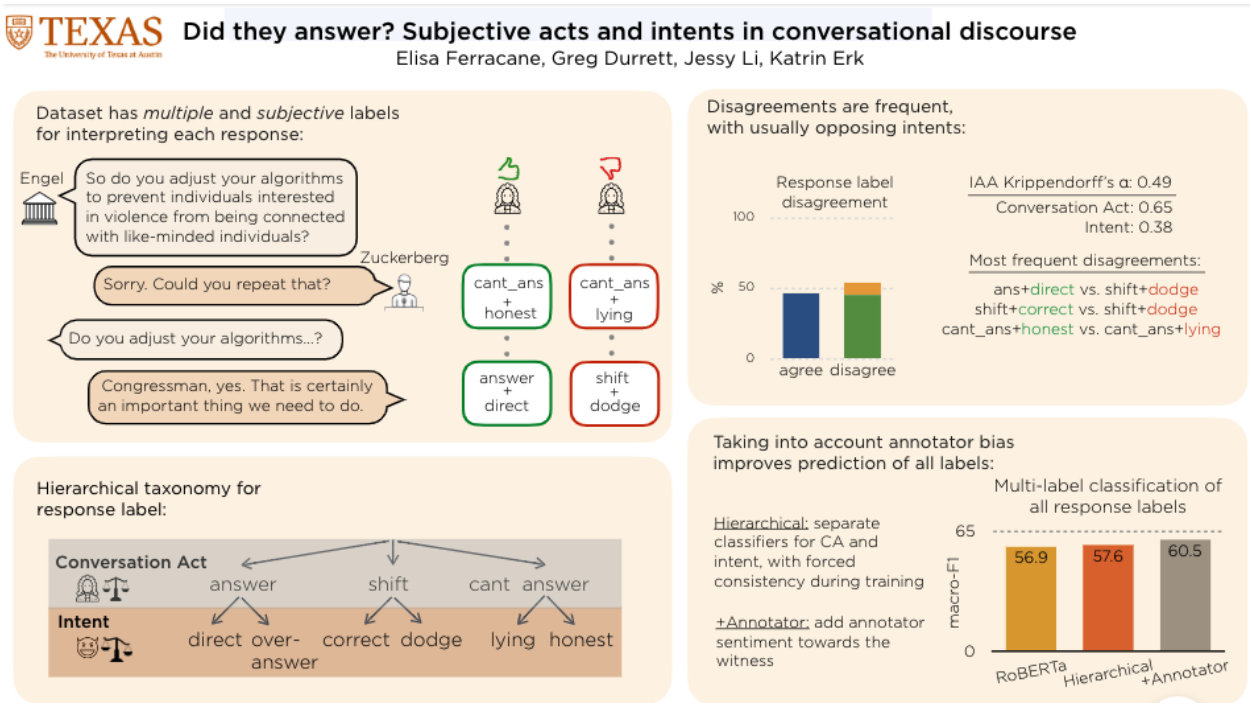
Papers in discourse & pragmatics

Session 5B

Bridging anaphora resolution: making sense of the SOTA

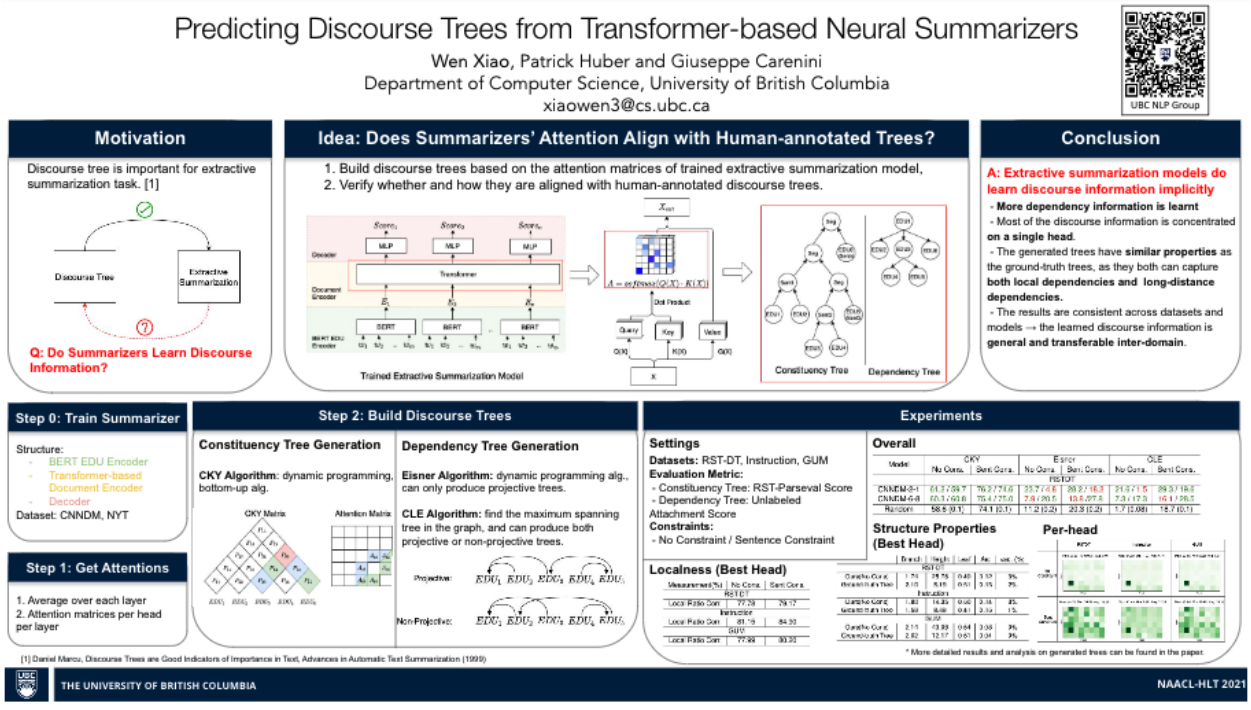


Did they answer? Subjective acts and intents in conversational discourse



Session 12A

Predicting discourse trees from Transformer-based neural summarizers



Is incoherence surprising? Targeted evaluation of coherence prediction from LMs

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* PERSONACHAT *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)

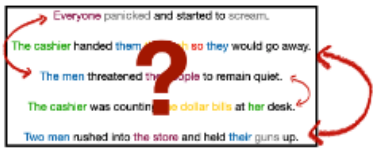


Figure 3: *Which notions of coherence does random shuffling break exactly?*

Targeted Syntactic Evaluation (Marvin & Linzen 2018)

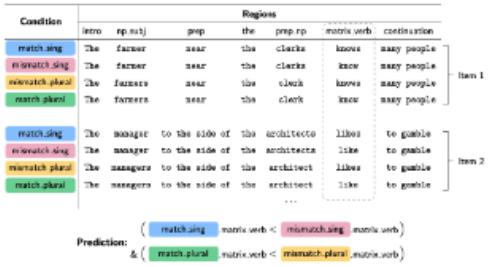


Figure 4: *SyntaxGym (Gauthier et al. 2020): Framework for syntactic evaluation of language models using minimal pairs and predictions evaluated on model's surprisal scores*

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

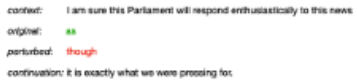
Pre-trained Models

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

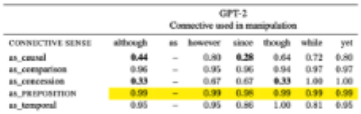
- Shown to perform well on down-stream tasks that require some notion of coherence, such as story generation (See et al. 2019), but automatic measures only suited to evaluate diversity, better methods needed to actually measure **text coherence**

Explicit Connectives

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



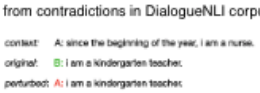
Results:



	GPT-2 Connective used in manipulation						
CONNECTIVE SENSE	although	as	however	since	though	while	yet
as_causal	**0.44**	–	0.80	**0.28**	0.64	0.72	0.80
as_comparison	0.96	–	0.95	0.96	0.94	0.97	0.97
as_concession	**0.33**	–	0.67	0.67	**0.33**	1.00	1.00
as_PREPOSITION	0.99	–	0.99	0.98	0.99	0.99	0.99
as_temporal	0.95	–	0.95	0.86	1.00	0.81	0.95

Speaker Commitment

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



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 Pandit[1] and Yufang Hou[2]

[1]INRIA, Lille, France [onkar.pandit@inria.fr] [2]IBM 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.

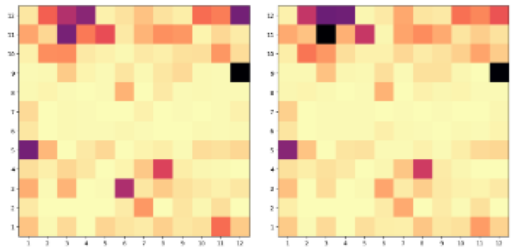


Figure 1:Bridging signals with BERT-base-cased model where anaphor and antecedents are 2 sentences apart. Bridging signals from anaphor to antecedent are shown in the first heatmap and the reverse signals in the second. In both heatmaps, the x-axis shows the attention head number and the y-axis shows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

Observations:

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

Of-cloze task investigation

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

Experimental Set-up:

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
Prominent attention heads		
(1) Salient/nearby mentions	20.15	-
Of-Cloze Test		
(2) Salient/nearby mentions	31.64	33.71
(3) All previous mentions	26.36	28.78
Of-Cloze Test: Ante. in the provided contexts		
(4) All previous mentions	29.00	30.88
Of-Cloze Test: Ante. out of the contexts		
(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.

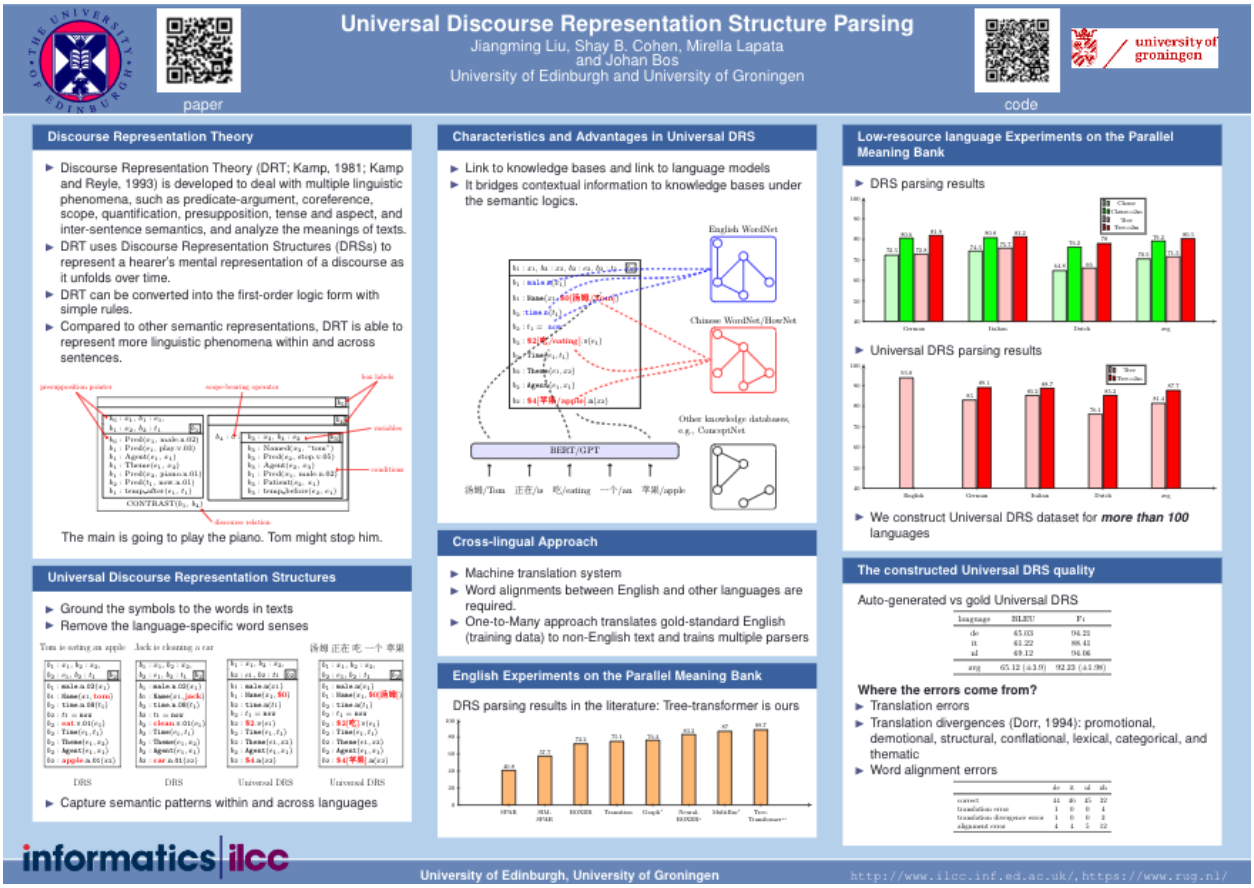
Distance	Accuracy
salient*	38.65
0	26.92
1	20.58
2	17.30
>2	10.98

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

References

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

Universal discourse representation structure parsing



Decontextualization: making sentences stand-alone

Decontextualization: Making Sentences Stand Alone

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

Data / Demo at GitHub!

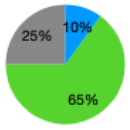
Overview

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

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

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 Unnecssary

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!

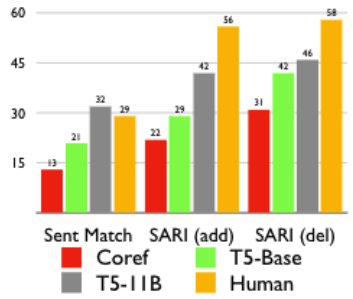
Intrinsic Evaluation

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:

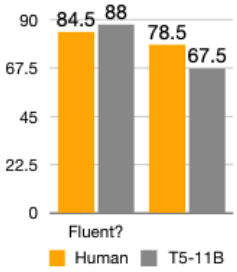
Base line	Coref	SpanBERT model, Untrained baseline
Base line	BART-small	Seq2Seq model, fine-tuned with decontextualized data
Base line	BART-large	Seq2Seq model, fine-tuned with decontextualized data
Upper Bounds	Human	Annotator

Results



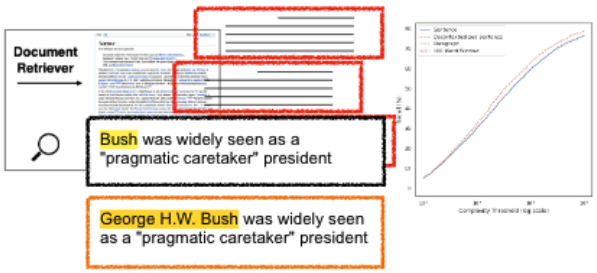
- Well trained seq2seq model performs surprisingly well
- Vanilla evaluation metric is challenging to tell apart good vs. bad decontextualization

Manual Eval



Application 1: Decontextualization as preprocessing

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



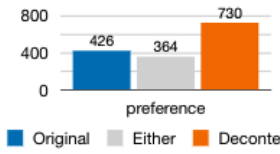
Application 2: Decontextualization As Is

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

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

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

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

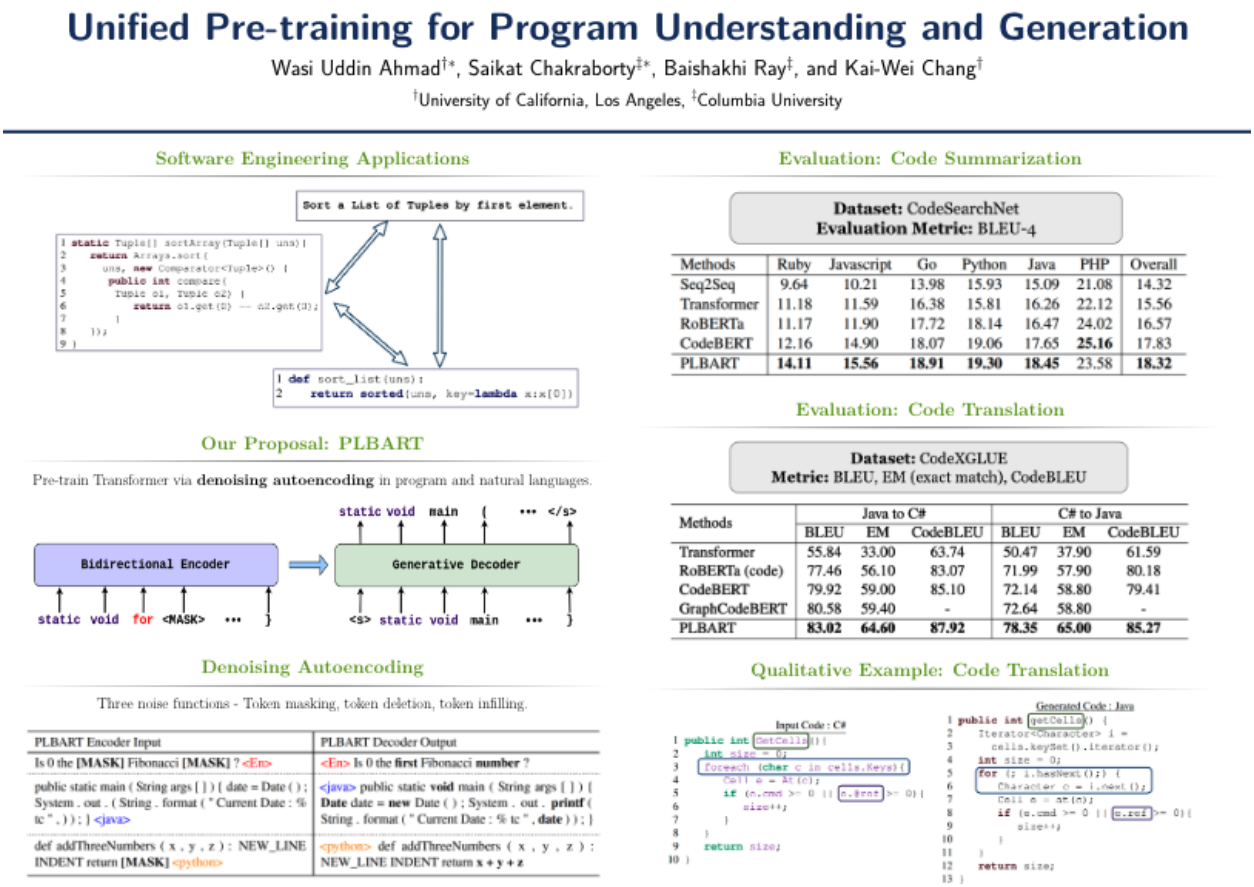


Users preferred decontextualized answer sentence, for providing succinct yet informative answer to the question.

Papers in ML4NLP

Session 8A

Unified pre-training for program understanding and generation



Unified Pre-training for Program Understanding and Generation

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

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

Software Engineering Applications

Sort a List of Tuples by first element.

```
static Tuple[] sortArray(Tuple[] uns){
  return Arrays.sort(
    uns, new Comparator<Tuple>() {
     public int compare(
      Tuple o1, Tuple o2) {
        return o1.get(0) - o2.get(0);
      }
  });
}
```

```
def sort_list(uns):
    return sorted(uns, key=lambda x:x[0])
```

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]** ? <En>	<En> Is 0 the **first** Fibonacci **number** ?
public static main (String args [ ]) { date = Date ( ) ; System . out . (String . format (" Current Date : % tc " ,)) ; } <java>	<java> public static **void** main (String args [ ]) { **Date** date = **new** Date ( ) ; System . out . **printf** (String . format (" Current Date : % tc " , **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#

```
public int GetCells(){
  int size = 0;
  foreach (char c in cells.Keys){
    Cell o = At(c);
    if (o.cmd >= 0 || o.ref >= 0){
      size++;
    }
  }
  return size;
}
```

Generated Code : Java

```
public int getCells() {
  Iterator<Character> i =
    cells.keySet().iterator();
  int size = 0;
  for (; i.hasNext();) {
    Character c = i.next();
    Cell o = At(c);
    if (o.cmd >= 0 || o.ref >= 0){
      size++;
    }
  }
  return size;
}
```

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**.

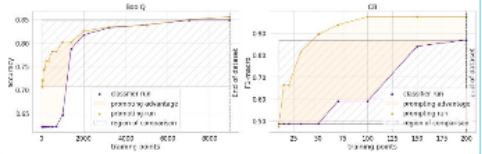
In contrast, a prompted model is used as a token predictor. It is presented with the input of the task and a *prompt:* **a short descrition of the task** that needs to be completed by a masked token. That token is then mapped into the desired class output by a *verbalizer*. (for example, *Yes* for 1 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 (0 Training Points)								
	MNLI	BoolQ	CB	COPA	MultiRC*	RTE	WiC	WSC
P vs H	3506 ± 536	752 ± 46	90 ± 2	288 ± 242	384 ± 378	282 ± 34	424 ± 74	281 ± 347
P vs N	150 ± 252	299 ± 81	78 ± 2	-	74 ± 56	404 ± 68	-354 ± 166	-
N vs H	3355 ± 612	453 ± 90	12 ± 1	-	309 ± 320	-122 ± 62	-70 ± 160	-

4/ Zero-shot vs adaptation

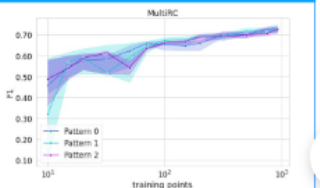
In order to study the zero-shot vs. adaptive nature of prompts, we introduce *null verbalizers*: models whose verbalizers are replaced with first names, so that zero-shot capability is at random chance. We compute which part of the data advantage is due to the zero-shot capability (*prompt* vs *null*)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

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

MultiRC*	RTE	WiC
384 ± 378	282 ± 34	424 ± 74
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

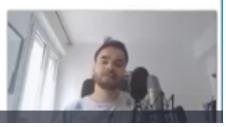
Testing on SuperGLUE + MNLI
For every task, we fine-tune models on subsets of increasing data sizes
Best of 4 runs on every data size

Linear head model

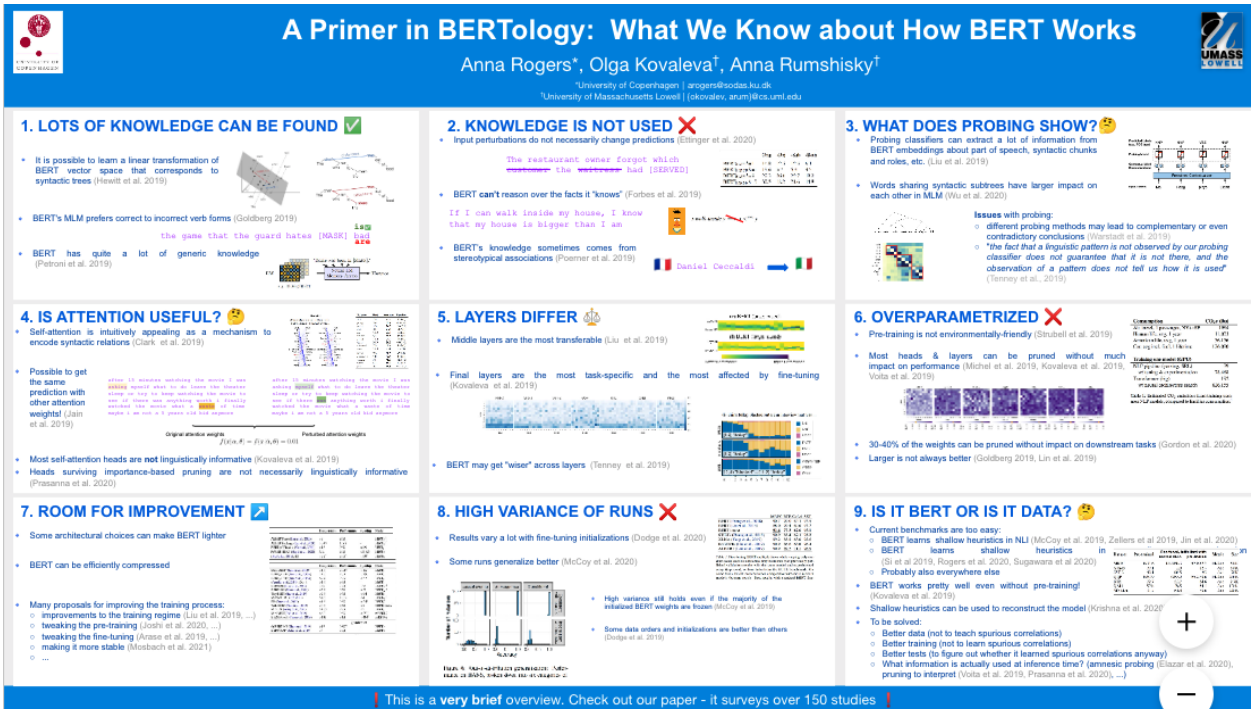
- Start from RoBERTa-large
- Linear classification head instead of prediction head
- Fine-tuned via backpropagation on the predicted class
- Slight hyper-parameter tuning to be within 2 points of SuperGLUE leaderboard

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.

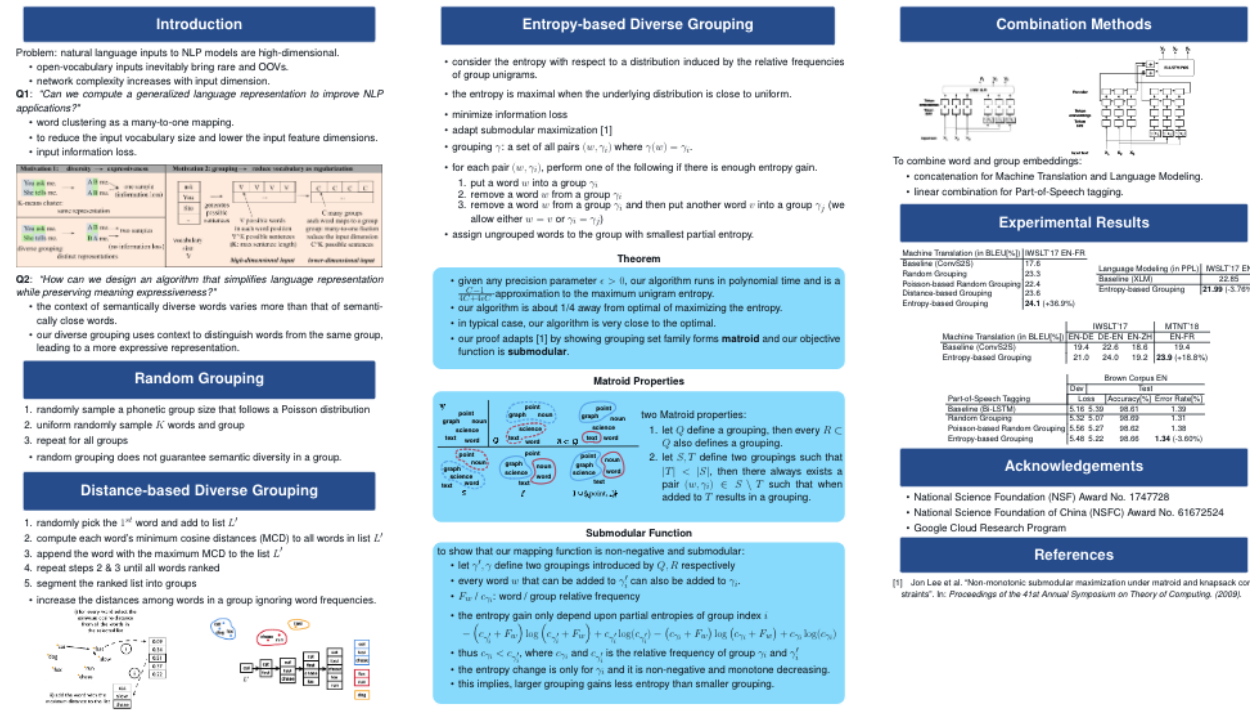


A primer in BERTology: What we know about how BERT works?

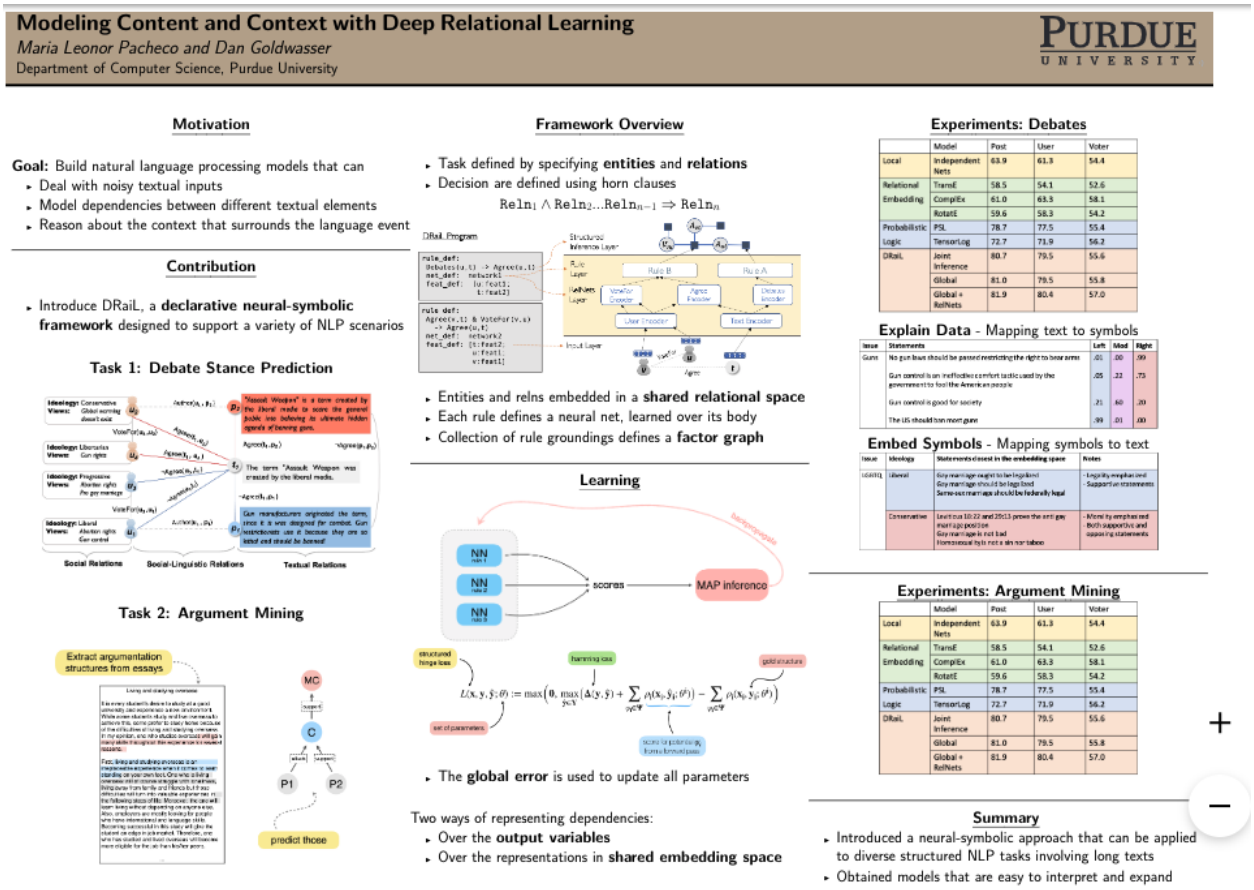


Session 9E

Grouping words with semantic diversity



Modeling content and context with deep relational learning



Session 14D

Revisiting simple neural probabilistic LMs

UMassAmherst
College of Information & Computer Sciences

Revisiting Simple Neural Probabilistic Language Models

Simeng Sun, Mohit Iyyer

https://github.com/SimengSun/revisit-nplm

KEY TAKEAWAYS

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

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

Two Transformer-based variants

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

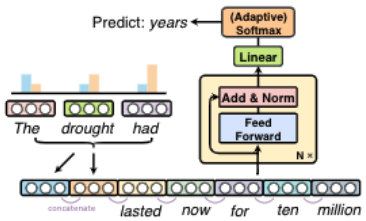
ABLATION

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

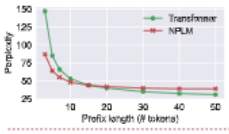
- Three word-level benchmarks (WIKITEXT-2, WIKITEXT-103, LAMBADA) and one character-level benchmark (ENWIK8).

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*♯, Aaron Jaech♭, Xin Li♯, Matt Gormley♯, Jason Eisner♯
*kitsing@cs.jhu.edu
♯Johns Hopkins University
♭Facebook AI
Carnegie Mellon University

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

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

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

Efficiently Computable (EC):
an abstraction of **Energy-Based Models (EBMs).**
A normalizable efficiently computable weighted language defines $p(\mathbf{x}) \propto \bar{p}(\mathbf{x})$ where $\bar{p}(\mathbf{x})$ can be computed in $O(\text{poly}(|\mathbf{x}|))$.
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 $\mathbf{x}$ as $p(\mathbf{x}) = \prod_t p(x_t \mid \mathbf{x}_{<t})$ with a fixed size parameter vector. Computing $p(x_t \mid \mathbf{x}_{<t})$ takes $O(\text{poly}(t))$.

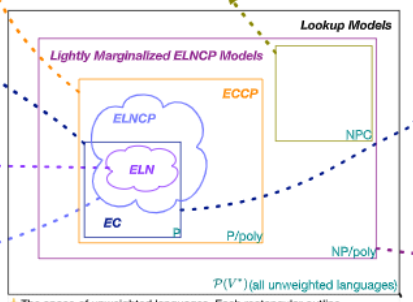
Efficiently Locally Normalizable with Compact Parameters (ELNCP):
a generalization of ELNs. An ELNCP model has infinitely many parameter vectors. When $|\mathbf{x}| = n$, an ELCNP model uses parameters θ_n to compute $p(\mathbf{x}) = \prod_t p(x_t \mid \mathbf{x}_{<t})$. They provide a conceptual upper 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 $p(\mathbf{x}\#\mathbf{y}) > 0$ iff $\mathbf{y}$ is a valid continuation of prompt $\mathbf{x}$, 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
$p(\underbrace{\mathbf{x}}_{\text{problem encoding}} \# \underbrace{\mathbf{y}}_{\text{solution candidate}})$
where $p(\mathbf{x}\#\mathbf{y}) \neq 0$ iff $\mathbf{y}$ is a solution to $\mathbf{x}$.
The **prefix probability** $p(\mathbf{x}\#) > 0$ if and only if $\mathbf{x}$ has a solution.
In other words, autoregressive models that factor $p(\mathbf{x}\#\mathbf{y}) = p(\mathbf{x}\#) \cdot p(\mathbf{y} \mid \mathbf{x}\#)$ **must** have the capacity to decide whether $\mathbf{x}$ has a solution, to ensure the joint distribution is accurate. If $\mathbf{x}$ 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/ELNCPs)!
The other sequence order does fine under autoregressive models (if $\mathbf{x}$ is in NP):
$p(\underbrace{\mathbf{y}}_{\text{solution candidate}} \# \underbrace{\mathbf{x}}_{\text{problem encoding}}) = p(\mathbf{y}\#) \cdot p(\mathbf{x} \mid \mathbf{y}\#)$
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 $p(x_t \mid \mathbf{x}_{<t})$!
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 des 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:
$p(\underbrace{\mathbf{x}}_{\text{problem encoding}} \# \underbrace{\mathbf{y}}_{\text{solution candidate}}) = p(\underbrace{\mathbf{y}}_{\text{latent variable}} \#) \cdot p(\underbrace{\mathbf{x}}_{\text{problem encoding}} \mid \underbrace{\mathbf{y}}_{\text{latent variable}} \#)$
They can have *any* language in NP/Poly as support!
$p(\underbrace{\mathbf{x}}_{\text{problem encoding}}) = \sum_{\mathbf{y} \in V^*} p(\underbrace{\mathbf{y}}_{\text{latent variable}} \#) \cdot p(\underbrace{\mathbf{x}}_{\text{problem encoding}} \mid \underbrace{\mathbf{y}}_{\text{latent variable}} \#)$
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.

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

Stanford NLP

code release@github: tatsu-lab/mlm_inductive_bias

1 Problem Statement

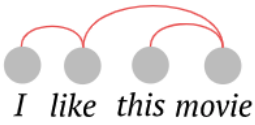
task-agnostic masking

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

2.1 Cloze Reduction Hypothesis

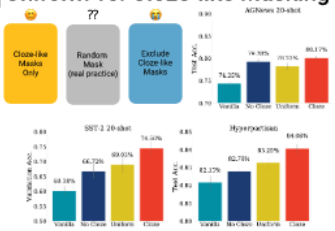
Cloze-like masking can provide indirect supervision but in practice we apply random masking.
We quantify the importance of cloze-like masking through controlled experiments.

2.2 Dependency Learning Hypothesis



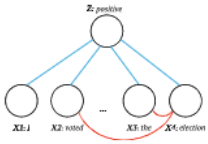
We hypothesize that generic masks can help learn the statistical dependencies among words and these dependencies are related to syntactic structures

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 $\Sigma_{\mathbf{XX}}$ is full rank,*

$$\mathbf{x}_{mask,i} = \beta_{2SLS,i}\mathbf{X}_{\setminus i} + O(\left\|\Sigma_{\mathbf{XX},\setminus i,i}\right\|_2),$$

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 $H(\mathbf{Z}\,X_{\setminus\{i,j\}})$.*

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

MLM objective directly ensures good approximationof cond. MI.

Proposition 4. *Let*

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

be an estimator constructed by the model distribution p_θ. Then we can show,

$$|\hat{I}_{p_\theta} - I_p| \leq \mathop{\mathbb{E}}_{x_j} D_{kl}\left(p(x_i|X_{\setminus i}(j, x_j))\|p_\theta(x_i|X_{\setminus i}(j, x_j))\right),$$

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	$\mathbf{58.74} \pm 0.22$

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