

NSNLI: First Workshop on Neuro-Symbolic methods for Natural Language Inference

Is Neuro-symbolic SOTA still a myth for Natural Language Inference

Somak Aditya, Microsoft Research
Maria Chang, IBM Research
Swarat Chaudhuri, UT Austin
Monojit Choudhury, Microsoft Research
Sebastijan Dumančić, KU Leuven



The Goal of the Workshop



Observation: A discrepancy in performance of large language models on benchmarks vs out-of-distribution simpler examples.



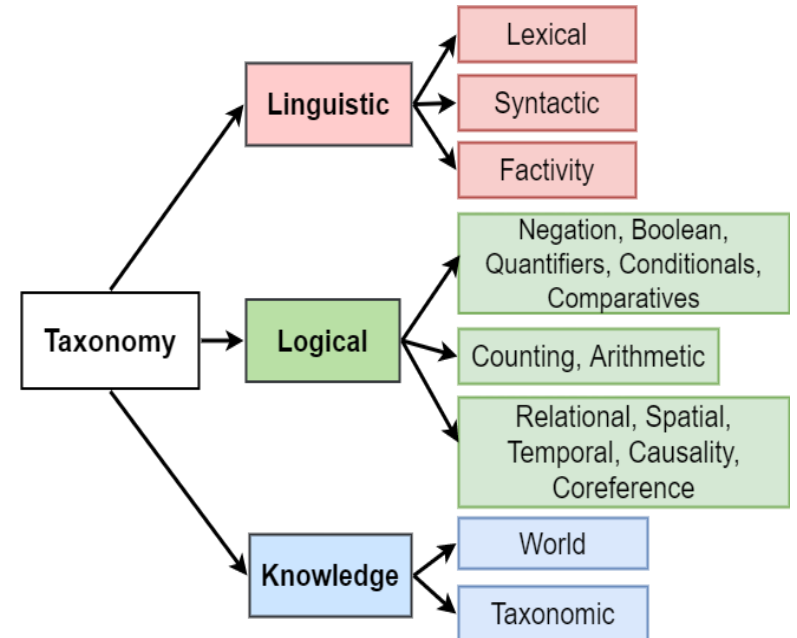
Lack of reasoning. Lack of generalization.



For reasoning, logic may help. *But how? What type of reasoning?*

The NLI Task (a stand-in for NLU)

Text	Hypothesis	Judgement
An older and younger man smiling.	The older man is not smiling	Contradiction
An older and younger man smiling.	Two men are laughing at the cats playing on the floor.	Neutral
An older and younger man smiling.	An older man is smiling.	Entailment



Task: Given a premise and hypothesis in natural language, identify whether the hypothesis contradicts, entailed by or neutral by the premise.

Critiques of NLI Systems

Jul 2018

Mar 2019

Feb 2020

	Breaking NLI	Fragments	Counterfactual NLI
Examples	P: The man is holding <i>saxophone</i> . H: The man is holding <i>electric guitar</i> .	P: Arthur visited Paris and New York H: Arthur did not visit Paris.	P: Students are inside of a lecture hall. OH: Students are indoors. (E) NH: Students are on the soccer field. (Contradiction)
Tests	Hypothesis and premise varies only one word.	Logical fragments: negation, Boolean, quantification.	Confounding factors: Entity, relations, actions. Semantic/Logical: Evidence, Negation.
Performance Drop	20% on LSTM based models	40% on Logic Fragments For Pre-trained BERT	30% on the New Set for Pre-trained BERT

Motivation: What Linguistic/logical phenomenon does the trained model capture? And observation: these model break easily.

The Reason Behind Such Discrepancy

ML/DL: Overfitting. Lack of generalization.

Logic, KR, Cognitive Science, and Linguistics:

- Lack of reasoning ability. (*Marcus 2018, 2020;*)
- Lack of commonsense knowledge. (*Davis & Marcus CACM 2015; Rao CACM 2021*)
- Lack of grounding and pragmatic abilities. (*Bender & Koller ACL 2020, Linzen ACL 2020*)

Which solution to Adapt?

At Least three Broad Camps

Deep Learning

Complete End-to-End NN Learning

Compile the symbolic knowledge (such as rule) to create data.

- Symbolic Math: Lample and Charton ICLR 2020

Mimic symbolic reasoning within a neural network

- Graph Neural Networks, Xu et.al. ICLR 2020

Probabilistic Logic/SRL

Cascade Neural and then Symbolic

Feed Neural outputs to symbolic reasoners.

No End-to-End Backprop

- Logic Tensor Networks. [PSL-VQA AAAI '18](#), and NSCL 2019 (MIT)

End-to-End Backprop

- Extension of Probabilistic Logic, (**DeepProbLog NeurIPS '19**, NLProlog ACL '19)

Program Synthesis + ATP

Algorithmic Inference with Neural Sub-routines

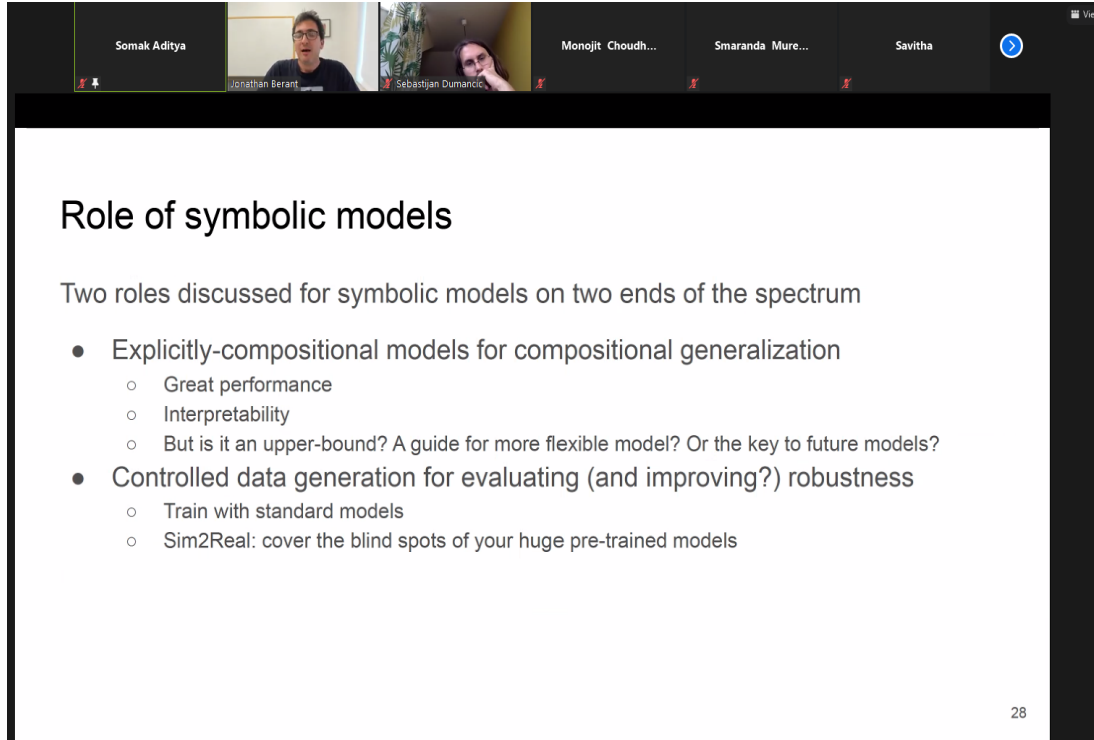
Symbolic systems makes the final decision. Neural is helping in steps.

- Example 1: ATP Provers (HOList)
- Example 2: Self-driving cars
- Example 3: Program synthesis+ML

Iterative Programs and ML

- Example 4: Program synthesis+ML

Session 1 – Jonathan Berant

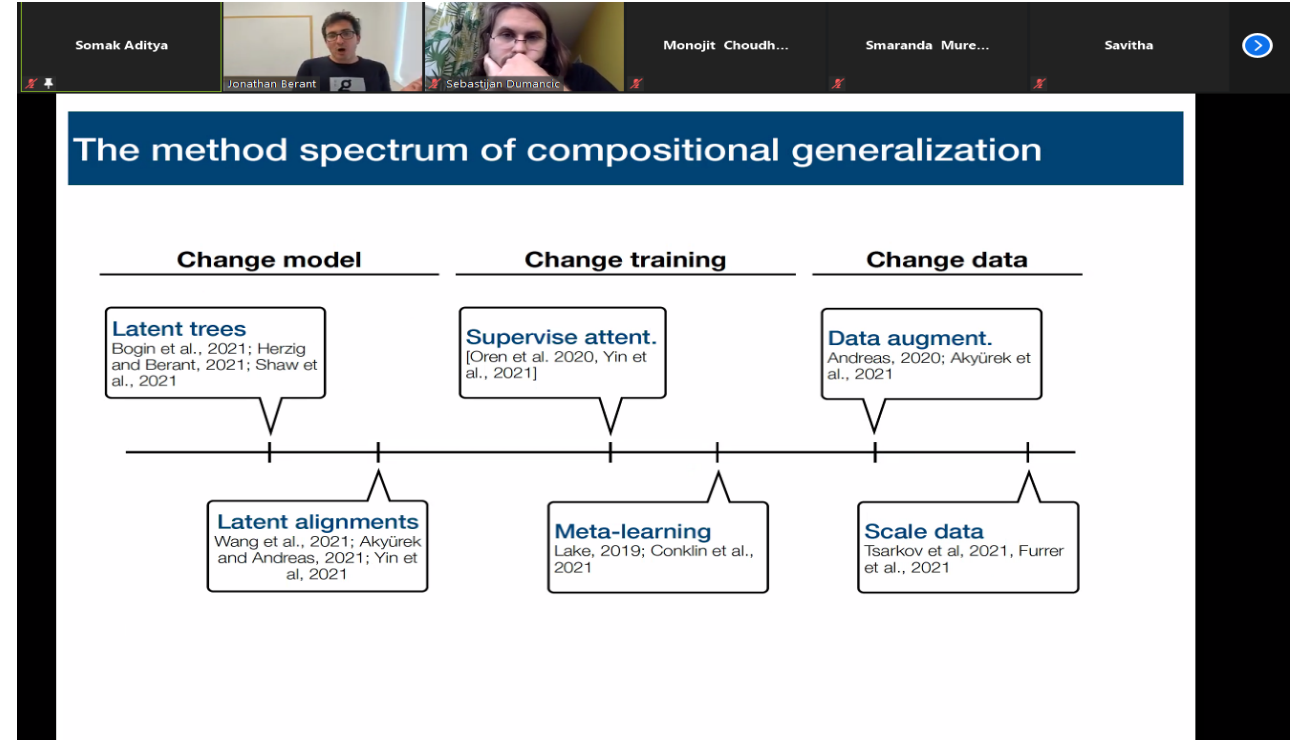


Role of symbolic models

Two roles discussed for symbolic models on two ends of the spectrum

- Explicitly-compositional models for compositional generalization
 - Great performance
 - Interpretability
 - But is it an upper-bound? A guide for more flexible model? Or the key to future models?
- Controlled data generation for evaluating (and improving?) robustness
 - Train with standard models
 - Sim2Real: cover the blind spots of your huge pre-trained models

28

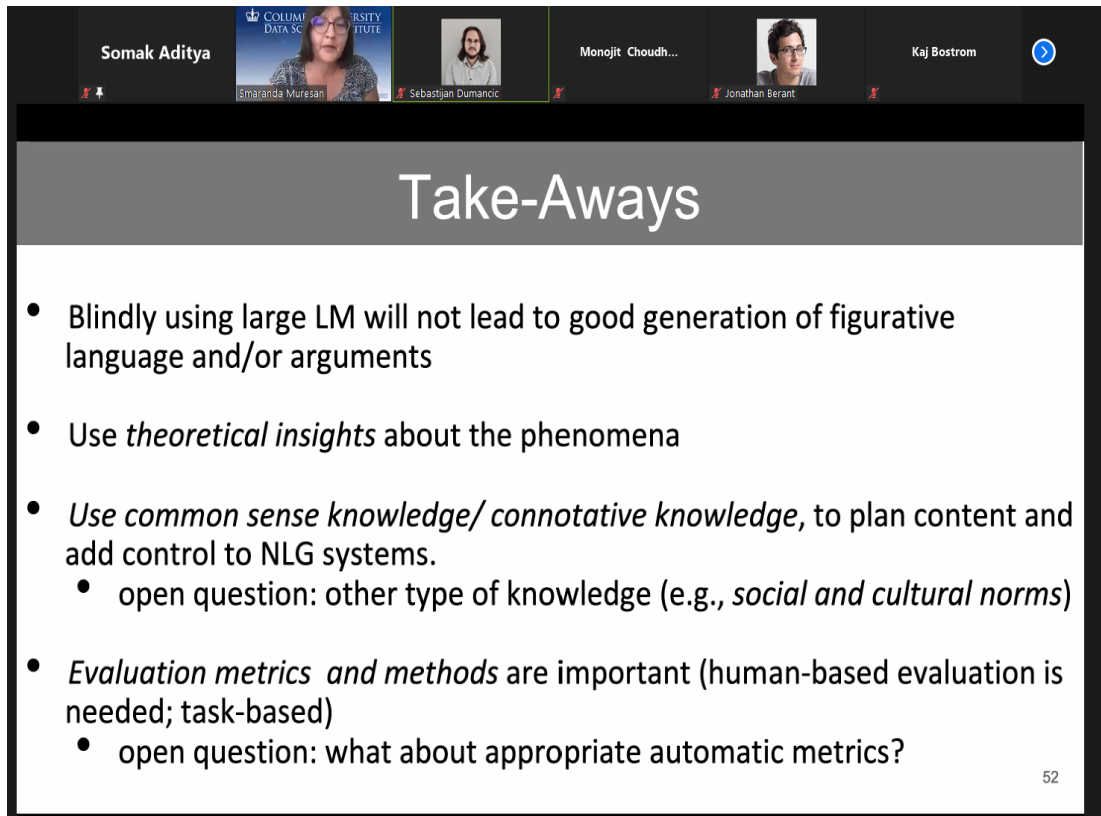


The method spectrum of compositional generalization

Change model	Change training	Change data
Latent trees Bogin et al., 2021; Herzig and Berant, 2021; Shaw et al., 2021	Supervise attent. [Oren et al. 2020, Yin et al., 2021]	Data augment. Andreas, 2020; Akyürek et al., 2021
Latent alignments Wang et al., 2021; Akyürek and Andreas, 2021; Yin et al., 2021	Meta-learning Lake, 2019; Conklin et al., 2021	Scale data Tsarkov et al., 2021, Furrer et al., 2021

Neuro-symbolic models for understanding complex questions*

Session 1 – Smaranda Muresan

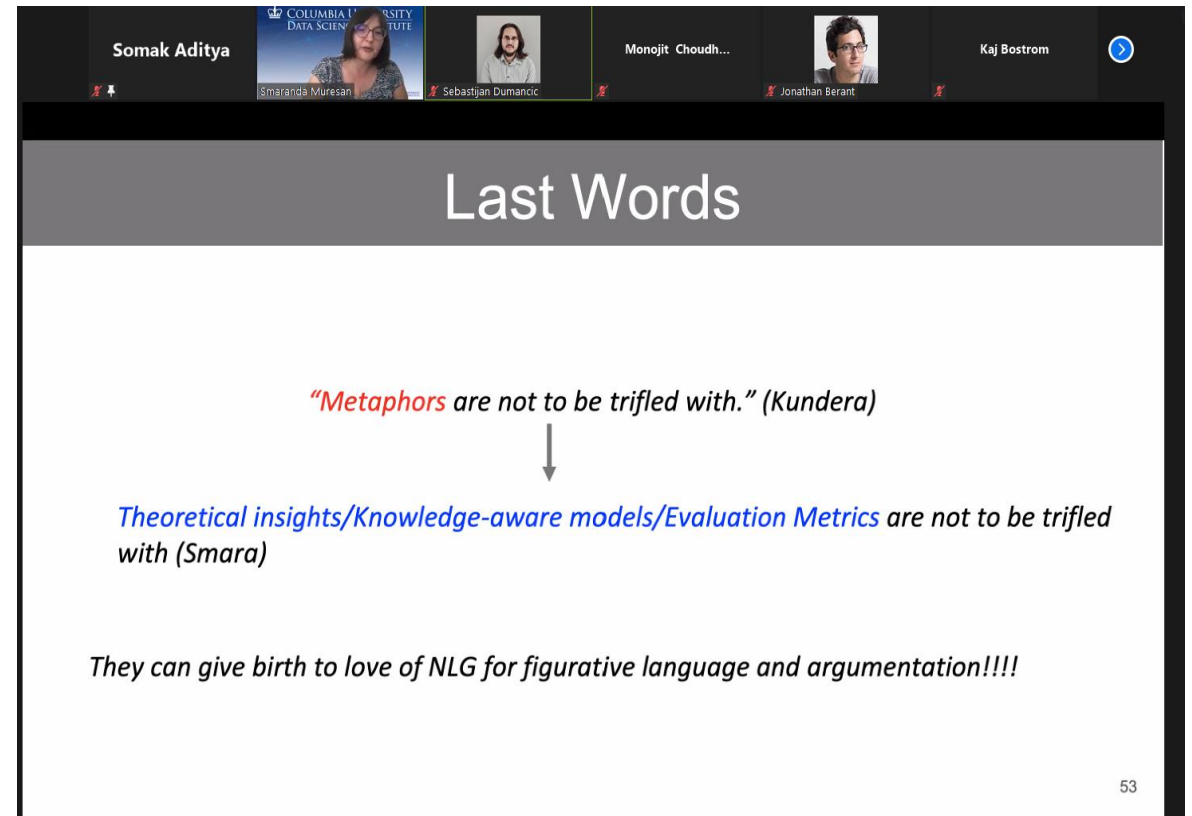


The screenshot shows a Zoom meeting interface with a slide titled "Take-Aways". The slide lists several bullet points regarding large language models and knowledge-enhanced text generation. The meeting participants visible at the top are Somak Aditya, Smaranda Muresan, Sebastian Dumancic, Monojit Choudhury, Jonathan Berant, and Kaj Bostrom.

Take-Aways

- Blindly using large LM will not lead to good generation of figurative language and/or arguments
- Use *theoretical insights* about the phenomena
- Use *common sense knowledge/ connotative knowledge*, to plan content and add control to NLG systems.
 - open question: other type of knowledge (e.g., *social and cultural norms*)
- *Evaluation metrics and methods* are important (human-based evaluation is needed; task-based)
 - open question: what about appropriate automatic metrics?

52



The screenshot shows a Zoom meeting interface with a slide titled "Last Words". The slide features a quote by Kundera, an arrow pointing to a paraphrase, and a concluding statement. The meeting participants visible at the top are Somak Aditya, Smaranda Muresan, Sebastian Dumancic, Monojit Choudhury, Jonathan Berant, and Kaj Bostrom.

Last Words

"Metaphors are not to be trifled with." (Kundera)

↓

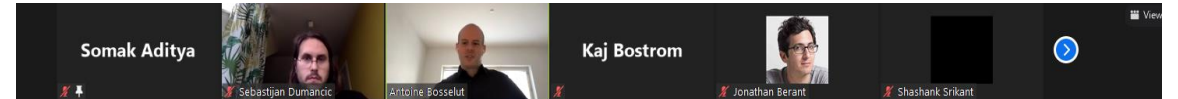
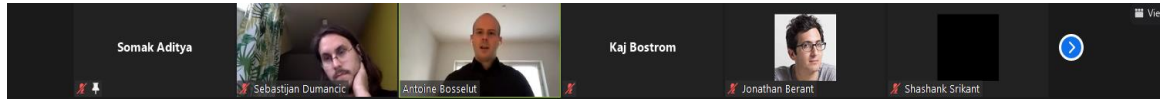
Theoretical insights/Knowledge-aware models/Evaluation Metrics are not to be trifled with (Smara)

They can give birth to love of NLG for figurative language and argumentation!!!!

53

Knowledge-enhanced Text Generation: The Curious Case of Figurative Language and Argumentation

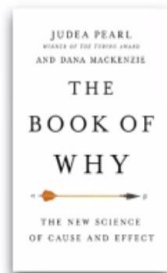
Session 1 – Antoine Bosselut



Reasoning with Deep Learning

- Deep learning models exploit **biases** (Bolukbasi et al., 2016), **annotation artifacts** (Gururangan et al., 2018), **surface patterns** (Li & Gauthier, 2017), etc.

They do not learn viable reasoning capabilities



(Pearl, 2018)

"All the impressive achievements of deep learning amount to just curve fitting"



Machines that Think like Humans

Understand situations by reasoning about commonsense knowledge

The trophy would not fit in the brown suitcase because **it was too big**. What was too big? (Levesque, 2011)



It's going to **snow**. I'll have to **wake up 30 minutes earlier**.



Symbolic Scaffolds for Neural Commonsense Representation and Reasoning

Thanks to Session 1 Speakers

Keynote Speakers:

- Jonathan Berant – *compositional model, controlled data generation*
- Smaranda Muresan – *knowledge-aware models, evaluation metrics, data generation*
- Antoine Bosselut – *representation and reasoning (with structure and unstructured)*

Invited and Contributed Papers:

- Meriem Beloucif – *BERT and the probing tasks*
- Shashank Srikant – *multiple demand system and language center*
- Kaj Bostrom – *alluding to a generative proof tree*

The open- floor QA!

A Hybrid QA-cum-panel.

- Host: Dr. Monojit Choudhury, Principal Researcher, Microsoft Research
- Questionees: Keynote and Invited Speakers
- Questioners: Audience
- Theme:
 - 1) The perceived disconnect between the philosophy and techniques for commonsense reasoning, and
 - 2) the need for more robust evaluation paradigms

Session 2

Schedule for the second Half-day session on August 22 (6:00 am – 9:30 am UTC).

6:00 – 6:30 **Welcome & Session 1 Summary**

Invited and Contributed Talks

6:30 – 8:30 (30 min) ProLinguist: Program Synthesis for Linguistics and NLP [paper](#)
(30 min) Reasoning using DeepProbLog [paper](#)
(20 mins) Perception, Inference, and Memory: The Trinity of Machine Learning [paper](#)
(20 min) A Generative-Symbolic Model for Logical Reasoning in NLU [paper](#)
(20 min) Multi-hop Reasoning Analysis Based on The Bayesian Probability [paper](#)

8:30 – 8:50 **Break**

8:50 – 9:15 **Open-Floor Q&A**

9:15 – 9:30 **Workshop Closing Statement**

Thanks to Session 2 Speakers

Invited Papers:

- Partho Sarathi – *program synthesis to learnt phonetic rules*
- Robin Manhaeve – *DeepProblog and its applications in CLUTTR*

Contributed Papers:

- Adam Lindstrom - *memory to be separate, alluding to continual learning*
- Jidong Tan – *Exciting application of generative neuro-symbolic network*
- Yitian Li – *Multi-hop Reasoning, Using intermediate hops to evaluate BERT*

Topics covered by Speakers

Transformers

- A Generative-Symbolic Model for Logical Reasoning in NLU
- Exploring Multi-hop Reasoning Process in NLU from the View of Bayesian Probability
- How can BERT Understand High-level Semantics?
- Keynote 2

Generative

- A Generative-Symbolic Model for Logical Reasoning in NLU
- Keynote 3

Probabilistic Logic

- Reasoning using DeepProbLog

Program Synthesis

- ProLinguist: Program Synthesis for Linguistics and NLP
- Flexible Operations for Natural Language Deduction
- Keynote 1

Cognitive science/NeuroScience

- Can Cognitive Neuroscience inform Neuro-Symbolic Inference Models?
- Perception, Memory, and Inference: The Trinity of Machine Learning