The Case for Scalable, Data-Driven Theory: A Paradigm for Scientific Progress in NLP

Julian Michael
New York University
julianjm@nyu.edu

Abstract

I propose a paradigm for scientific progress in NLP centered around developing *scalable, data-driven theories* of linguistic structure. The idea is to collect data in tightly scoped, carefully defined ways which allow for exhaustive annotation of behavioral phenomena of interest, and then use machine learning to construct explanatory theories of these phenomena which can form building blocks for intelligible AI systems. After laying some conceptual groundwork, I describe several investigations into data-driven theories of shallow semantic structure using Question-Answer driven Semantic Role Labeling (QA-SRL), a schema for annotating verbal predicate–argument relations using highly constrained question-answer pairs. While this only scratches the surface of the complex language behaviors of interest in AI, I outline principles for data collection and theoretical modeling which can inform future scientific progress. This note summarizes and draws heavily on my PhD thesis (Michael, 2023).

1 Introduction

Formal representations of linguistic structure and meaning have long guided our understanding of how to build NLP systems, *e.g.*, in the traditional NLP pipeline (Jurafsky and Martin, 2008). However, this approach has always had limitations:

1. Fully specifying formal representations requires resolving challenging theoretical questions long contentious among linguists;
2. It is difficult to reliably produce these representations with broad coverage using machine learning; and,
3. Even ostensibly correct linguistic representations are often hard to apply downstream.

Together with the effectiveness of deep learning, these challenges led to the proliferation of end-to-end neural network models which directly perform tasks without intermediate formal representations of linguistic structure (He et al., 2017; Lee et al., 2017; Seo et al., 2017, *inter alia*). This trend continues with language model assistants like GPT-4 (OpenAI, 2023) and Claude (Bai et al., 2022) which can perform a wide range of tasks. However, these systems are still not robust, often reporting false or biased answers (Perez et al., 2022; Bang et al., 2023) and making false claims about their own reasoning (Turpin et al., 2023). Ensuring AI systems' robustness requires us to precisely characterize and control their generalization behaviors.

To this end, formal theories, *e.g.*, of linguistic structure, common sense, reasoning, and world knowledge, provide frameworks for evaluation. They inform the design and construction of challenge sets (McCoy et al., 2019; Naik et al., 2018; Wang et al., 2019), measures of systematicity (Yanaka et al., 2020; Kim and Linzen, 2020), behavioral tests (Linzen et al., 2016), and probing experiments (Liu et al., 2019; Tenney et al., 2019). As these theories allow us to characterize generalization behaviors we desire, they will likely play a pivotal role in the design and training of trustworthy systems. So core improvements in formal theories of aspects of intelligent behavior may yield boons for both the construction and evaluation of NLP systems. But the question remains of how to achieve this: decades of work on semantic ontologies (Baker et al., 1998; Palmer et al., 2005), commonsense knowledge bases (Lenat, 1995; Speer et al., 2017), and formal reasoning systems (Lifschitz, 2008) have largely been superseded in NLP by deep learning and language models.

Theory-driven approaches in AI have been so disappointing that Sutton (2019) famously argues that intelligence and the world are simply too complex for us to capture with domain theories, and we should instead focus on general-purpose learning systems that can capture this intrinsic complexity from data. However, I believe this is too pessimistic, giving up on the *intelligibility* of AI systems that is provided by accurate theories of their behavior, which is necessary for verifying their safety and usefulness in high-risk, high capability settings (Ngo et al., 2023). Instead, the deep learning era presents an opportunity to rethink how we develop theories of language behavior.

In particular, I propose *scalable, data-driven theory* as a paradigm to address the shortcomings mentioned at the beginning of this article: resolving or sidestepping theoretical questions, producing representations with broad coverage, and applying them effectively in downstream tasks. Inspired by Pragmatist epistemology (James, 1907), this approach avoids requiring the linguist or theoretician to specify the entire theory by hand, instead integrating machine learning in a judicious way which allows for the scalable, automated induction of formal theoretical constructs (*e.g.*, ontologies) which are grounded in task-relevant linguistic behaviors.

2 Pragmatist Principles for Scientific Progress

Church (2007) describes the history of computational linguistics on a *pendulum*, swinging between Rationalist (theory-driven) and Empiricist (data-driven) paradigms every 20 years. Church lists the "swings" as follows (with my comments):

- 1950s: Empiricism (Shannon, Skinner, Firth, Harris) — information theory, psychological behaviorism, early corpus linguistics
- 1970s: Rationalism (Chomsky, Minsky) — generative linguistics, logic-based AI
- 1990s: Empiricism (IBM Speech Group, AT&T Bell Labs) — statistical NLP, machine learning, modern distributional semantics
- 2010s: A Return to Rationalism?

As the reader may know, the predicted "Return to Rationalism" did not happen. NLP, for its part, is more Empiricist than ever.

Why is this? Sutton may say it's because the world is too complex: The Rationalist theoretician carefully formalizing the problems at hand has no hope of capturing the world's intricacies in a manually-crafted theory, though a system implementing that theory can be understood and controlled. The Empiricist tinkerer, on the other hand, can build a system that mostly works by trial, error, patching and fastening; so they win on empirical benchmarks. However, the resulting system is too complex to fully understand or control, and generalizes in unpredictable ways.

An odd feature of the Rationalism/Empiricism dichotomy is that neither epistemology accurately describes the pursuit of science in most fields. In fields like physics, chemistry, and biology, theoretical and experimental approaches are not in conflict; rather, they synergize and inform each other, as theories are continually updated to align with new experimental data. To make sense of this, we can turn to an epistemology inspired by how people actually operate in the world: Pragmatism.

Pragmatism is an epistemological framework which conceptualizes *knowing* in terms of the *actions* that the knowledge licenses, *i.e.*, by the predictions that follow from that knowledge. Prominent Pragmatists include Charles Sanders Peirce (1839–1914) and William James (1842–1910). Like Empiricism, Pragmatism embraces experience as the primary source of knowledge. But unlike Empiricists, Pragmatists such as James embrace formal and linguistic categories as comprising the content of knowledge, on the basis of their *usefulness* in making predictions and licensing actions (James, 1907). Unlike in Rationalism, the Pragmatist search for truth is not a search for one true theory which fundamentally describes the world, but for an ever-expanding set of theoretical tools and concepts that can be picked up and put down according to the needs of the knower. In pithy terms, a Pragmatist might agree with the statistical aphorism that that "All models are wrong; some are useful" (Box, 1976). Pragmatists such as James (1907) claim that this perspective more accurately describes human behavior with respect to knowledge (and indeed, the pursuit of science) than prior epistemologies.

Combining the core ideology of Pragmatism with observations from computational linguistics, we can derive two guiding principles for the development of theories that may have prospective use in NLP: decouple data from theory (Section 2.1), and make data reflect use (Section 2.2).

2.1 Decouple Data from Theory

One feature that distinguishes much NLP work, particularly involving linguistic structure, from traditional sciences is the status of theory with respect to data. In most empirical sciences, data takes the form of concrete measurements of the world, and the task of a theory is to explain those measurements. In NLP, many benchmarks and datasets are constructed under the *assumption* of a theory, whether it be one of syntactic structure (Marcus et al., 1993; de Marneffe et al., 2021), semantic structure (Palmer et al., 2005; Banarescu et al., 2013), or some other task-specific labeling scheme.

A theory, *e.g.*, of syntactic or semantic structure, is useful for annotation in providing a straightforward way to annotate disambiguation of text, which is important for understanding language. However, errors and inconsistencies in annotation resulting from complexity, vagueness, or underspecification in the theory limit what can be learned by models, as human performance and inter-annotator agreement can be surprisingly low (Nangia and Bowman, 2019). For example, the OntoNotes compendium of semantic annotations (Hovy et al., 2006) was presented as "The 90% solution" because of 90% agreement rates — implying that the dataset cannot validate performance numbers higher than 90%.

As another example, Palmer et al. (2006) find that fine-grained sense distinctions produce considerable disagreement among annotators of English text. But fixing the problem can't just be a matter of improving the sense inventory: they find that coarser-grained sense groups designed to improve agreement lack the distinctions from fine-grained senses that are necessary for predicting how words should translate into typologically distant languages like Chinese and Korean. When different tasks require different theoretical distinctions, setting them in stone during annotation is a problem, especially considering that there will almost certainly be missing categories, as new word senses or distinctions may show up in more exhaustive data or under domain shift. More generally, refining annotation guidelines to increase agreement between annotators does not necessarily solve the problem, as the extra assumptions built into the annotation process do not necessarily encode any more scientifically meaningful information in the data — a problem known in the philosophy of science as the *problem of theoretical terms*.[1]

Building a robust theory that can scale to unexpected phenomena and new data, and be adjusted for new tasks, requires theoretical agility which is precluded by committing to a theory-based annotation standard. An alternative is to directly annotate the phenomena that the theory is meant to explain, and derive the theory on the basis of this data. This, for example, is how *grammar engineering* is done in the DELPH-IN consortium (Bender and Emerson, 2021). For each language, a broad-coverage Head-driven Phrase Structure Grammar (HPSG) is maintained separately from its associated treebank, which is annotated not with full syntactic analyses but with *discriminants* (Carter, 1997) such as prepositional phrase attachment sites which constrain the set of possible parses in a way that is independent of the grammar. Then, when the grammar is updated, the discriminants are used to automatically update the treebank while also providing data to validate the updated theory (Oepen et al., 2004; Flickinger et al., 2017). Pushing the envelope further are the Decompositional Semantics Initiative (White et al., 2016) and MegaAttitude project (White and Rawlins, 2016).[2] In these projects, annotating large-scale corpora with the phenomena that are posited to underly linguistic theories in question — such as Dowty (1991)'s proto-role properties, or entailments corresponding to neg-raising (An and White, 2020) and projection (White and Rawlins, 2018) — has facilitated insights regarding argument selection (Reisinger et al., 2015) and lexically-specified syntactic subcategorization rules (White, 2021), as well as automatically inducing lexicon-level ontologies of semantic roles (White et al., 2017) and event structure (Gantt et al., 2021) that are derived directly from the phenomena they are designed to explain.

The lesson of Empiricism is that for a model to work, it must be learned from data; while Rationalism tells us that for a model to be intelligible and general, it must be grounded in theory. A wealth of innovative prior work shows us that Pragmatism is possible: we can have both.

2.2 Make Data Reflect Use

A satisfying data-driven theory of a few linguistic phenomena is not sufficient as a backbone for general language understanding systems. The second relevant lesson of Pragmatism is that the model must be fit to its use. The approaches reviewed in Section 2.1 are, by and large, targeted at theoretical questions in language syntax and semantics, *e.g.*, regarding the nature of syntactic structure across many languages (Bender et al., 2002) or the syntactic realization of a verb's arguments (Reisinger et al., 2015). On the other hand, general-purpose

[1] See Riezler (2014) for a discussion of this issue in NLP.

[2] `https://decomp.io`, `https://megaattitude.io`

language processing relies on a huge amount of lexical and world knowledge and inferential ability which is outside the scope of traditional linguistic theories. While general-purpose syntactic and semantic representations have some direct uses in NLP end-tasks, such as for search and retrieval (Schäfer et al., 2011; Shlain et al., 2020), their application in downstream tasks requiring higher-level reasoning or inference, like reading comprehension, translation, and information extraction has been less fruitful. This is at least in part because these theories are far insufficient to serve as mechanistic accounts of the inferential phenomena which are required to perform those tasks.

Constructing theories which *can* account for such phenomena is a monumental challenge. But it is a challenge which, I argue, we must address if we want to pursue the goal of accurate, reliable, and intelligible systems. Pragmatism tells us the first step is to catalog the phenomena we wish to explain in a way that is amenable to theoretical modeling. This will require carefully carving up the space of phenomena in such a way that useful abstractions can be designed to facilitate future progress (Dijkstra, 1974); Section 4 will discuss considerations on how to do this well.

3 Scalable, Data-Driven Theory

The principles in Section 2 imply a general framework for building useful theories, which I call *data-driven theory*: First, annotate data in a theoretically-minimal way, scoped carefully to reflect specific phenomena that we want to explain; then, automatically induce theories to explain those phenomena using computational methods like machine learning. But how does this method scale in practice? Even if the resulting theories are high-quality, requiring annotated data limits their scope to orders of magnitude less than what is leveraged by standard pretrained models (Brown et al., 2020; OpenAI, 2023; Bai et al., 2022).

Black-Box Data Simulators This is where black-box models may actually be able to help. Even if they are uninterpretable on their own, their high accuracy and data efficiency means they can be used as *data simulators*, generating phenomenological data — potentially at a level of granularity or exhaustivity unobtainable from humans — which can be fed into another, more interpretable algorithm to distill a theory from it. This is the approach we take in Michael and Zettlemoyer (2021), described in Section 5: We first train a black-box model to generate QA-SRL questions, where each role is labeled with only a single question in the training data. Then we decode full question *distributions* from this model, and induce an ontology of semantic roles by clustering arguments based on the overlap of their question distributions. While this work required a large training set of QA-SRL annotations (FitzGerald et al., 2018), it may now be possible to do such experiments without large-scale human data annotation at all, thanks to recent advances in instruction following by language models (OpenAI, 2023; Bai et al., 2022).

It may seem like the use of a black-box model as a data simulator begs the question: if our concern is that the black-box model isn't learning the underlying function we hope it is, then doesn't using it to simulate data risk leading us to a theory of the wrong function? Well, yes — *but the theory lets us do something about it*. Examining the "wrong" parts of the resulting theory (*e.g.*, induced semantic roles that don't match what we intuitively expect, or that lead to downstream predictions we think are wrong), and their connection to the training data, will identify one of the following:

- Systematic gaps in the data or mistakes in the model used for data simulation — which can then be filled or corrected.
- Mistakes in the modeling assumptions used in the theory induction algorithm — giving us information useful for improving our theories.
- Mistakes in our intuition about what the theory should have looked like in the first place — which means we've learned something.

All of these are positive outcomes for scientific progress. See Michael and Zettlemoyer (2021) for an in-depth analysis of this kind.

Scaling in Complexity Even if we can scale a theory's *size*, *e.g.*, to a large knowledge base or linguistic ontology, this does not handle the case of more *complex* tasks, with more nuanced relations between input and output (such as open-ended question answering or common sense inference tasks). Since theoretical modeling requires narrowly-scoped data (discussed more in Section 4), I do not expect that we can construct theories of such broad capabilities in the short term. However, if we carve up the space of tasks to start with theories of simple sub-phenomena of reading and inference, then we may be able to bootstrap from these theories to annotate and make sense of more complex data — for example, one can imagine eventually inducing rich, broad-coverage entailment graphs in the style of Berant et al. (2015) or McKenna et al. (2023) on the basis of comprehensive annotations of structured inferences in context. A complete or "true" theory of complex NLP tasks may be impossible even in principle, but — in the spirit of Pragmatism — that doesn't mean we can't construct theories that are *useful* for understanding and controlling AI systems. How my proposed framework scales with task complexity is unclear as of yet, but scalable theories of narrow phenomena provide a step in the right direction.

4 Data: Scoping Language Behaviors

The first step to developing theories of linguistic structure in an empirical, data-driven way is to carefully choose the data. To guide this, I propose **Four Principles of Scientific Data for NLP**:

1. **Theoretical minimalism.** The data should rely on as few theoretical assumptions as possible. For example, to capture natural language syntax, you should directly annotate the *phenomena* that you intend your syntactic theory to explain rather than directly annotating theoretical constructs like syntactic trees. This creates the space for an underlying theory to meaningfully explain this data.

2. **Broad comprehensibility.** To facilitate on-demand data collection at large scale in new domains, it should be possible and affordable to recruit non-expert annotators to label large amounts of data (*e.g.*, through crowdsourcing), or it should be feasible to automatically generate the data (*e.g.*, with language models).

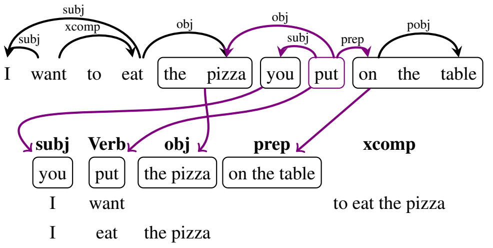
3. **Annotation constraints.** The output space of the task should be sufficiently constrained to allow for exhaustive coverage of the phenomena of interest. A task which is too open-ended leads annotators to produce a convenience sample of the output space, resulting in biased data that doesn't capture the full complexity of the phenomena of interest (Cai et al., 2017; Gururangan et al., 2018).

4. **Narrow scope.** The task should not capture too much complexity in the relationship between input and output. Not only can this make it difficult for annotators to reliably produce high-quality data, but it makes it more difficult to model the phenomena expressed in the data with a comprehensible theory.

Principles 1 and 2 instantiate Section 2.1's recommendation to decouple data from theory, while Principles 2, 3 and 4 help make it tractable to develop broad-coverage, comprehensible theories from this data. The final requirement is that the data reflect relevant downstream use cases (Section 2.2), which in our case means it should encode phenomena representing the intended behavior of AI systems performing language tasks.[3] I focus on a key strategy to meet these requirements: *annotating natural language with natural language question-answer pairs*. Question answering has long been used as a general-purpose format for testing language comprehension or executing practical language tasks (Gardner et al., 2019b; McCann et al., 2018; McCarthy, 1976), as nearly any task can be phrased as a question and questions which test a reader's comprehension of a text need not require specialized linguistic or theoretical expertise to answer. The downside of this great generality is that data annotation tends to be highly under-constrained and unsystematic (Gardner et al., 2019a), so we must judiciously constrain the space of question-answer pairs we use in accordance with the Four Principles.

This work is focused on annotations of shallow semantic structure: syntax, semantic roles, and other predicate–argument structure relations expressed in text. He et al. (2015) pioneered the use of question-answer pairs as a proxy for such structure in *Question-Answer driven Semantic Role Labeling* (QA-SRL), a framework for annotating English verbal predicate–argument relations using simple, highly constrained question-answer pairs. In the rest of this section, I will describe three data annotation projects which explored variations of this approach, illustrating some of the basic tensions between the Four Principles.

[3] This work is concerned with normative theories of AI behavior when performing language tasks. Insofar as we wish to produce theories of AI behavior which are comprehensible to us, aligned with our intuitions, and allow us to interface fluidly with machines using language, this goal should mostly be aligned with developing *descriptive* theories of *human* language behavior, which can then be used to constrain and guide AI behavior. The relationship between these theories and their importance for interacting with machines are discussed more in Chapter 2 of Michael (2023).



Dependency	Question	Answer
want → I	What wants to eat something?	I
eat → I	What would eat something?	I
eat → pizza	What would something eat?	the pizza
put → you	What put something?	you
put → pizza	What did something put?	the pizza
on → table	What did something put something on?	the table

Figure 1: Question-answer pair generation for human-in-the-loop parsing (He et al., 2016). We use the predicted CCG category of each verb to generate the questions, which are in in one-to-one relation with syntactic dependencies in the sentence. This one-to-one assumption was ultimately too strong, as workers answer these questions according to semantics and not just syntax.

4.1 Human-in-the-Loop Parsing

He et al. (2016) introduces *human-in-the-loop parsing*. We construct multiple-choice questions from syntactic attachment ambiguities in a parser's n-best list, get crowdsourced workers to answer these questions, and then re-parse the original sentence with constraints derived from the results (Figure 1). Testing on the English CCGbank (Hockenmaier and Steedman, 2007), we find only a small improvement in parser performance. A core challenge is the *syntax–semantics mismatch*, where workers provide answers which are semantically correct but correspond to the wrong syntactic attachment. For example, in the sentence "Kalipharma is a New Jersey–based pharmaceuticals concern that sells products under the Purepac label", workers unanimously answer the question "What sells something?" with "Kalipharma", which is not the syntactic subject of *sells* but a more natural way of referring to the same entity. So even though our annotation task is tightly scoped, our interpretation of the results requires theoretical assumptions which do not match the intuitions of non-expert workers.

4.2 Crowdsourcing Question-Answer Meaning Representations

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Who will **join** as **nonexecutive director**? - Pierre Vinken
What is **Pierre**'s last name? - Vinken
Who is **61 years old**? - Pierre Vinken
How **old** is **Pierre Vinken**? - 61 years old
What will he **join**? - the board
What will he **join the board** as? - nonexecutive director
What type of **director** will **Vinken** be? - nonexecutive
What day will **Vinken join the board**? - Nov. 29

Figure 2: Example Question-Answer Meaning Representation (Michael et al., 2018). Non-stopwords drawn from the source sentence are in bold. QAMR question–answer pairs capture a wide variety of relations, but are unstructured and hard to use downstream without extra tools such as a syntactic parser — here, our annotation task was too unconstrained and task scope too broad.

Michael et al. (2018) takes the opposite tack, broadening the task's scope by gathering open-ended questions from annotators to capture as many semantic relationships as possible in the source sentence. This requires adding many careful constraints and incentives to the crowdsourcing procedure, but we are careful to allow for open-ended questions that express annotator creativity. The result is a dataset of *Question-Answer Meaning Representation* (QAMR) annotations over English encyclopedic and news text covering many interesting phenomena (see Figure 2). However, achieving high recall of predicate–argument relations is not economical, requiring high annotation redundancy, and the unstructured question-answer pairs are hard to use downstream. The most successful use of QAMR in follow-up work is probably Stanovsky et al. (2018), where we convert QAMRs into Open Information Extraction tuples, but have to run the questions through a syntactic parser to do so. The lesson from these results is that leaving the annotation space too open and unconstrained leads to difficulties with recall and challenges with downstream modeling and theory.

4.3 Large-Scale QA-SRL Parsing

FitzGerald et al. (2018) returns to QA-SRL. In the original QA-SRL work (He et al., 2015), trained annotators specify the questions using drop-down menus in an excel spreadsheet. In this work, we streamline and scale up data collection, gathering high-coverage annotations for over 64,000 sentences with a two-stage generate/validate crowdsourcing pipeline (see Table 1 for examples). We increase annotation speed, reliability, and coverage using an autocomplete system which tracks the syntactic structure of QA-SRL questions as the annotator types, using it to suggest completions as well as whole questions. In terms of semantic richness and annotation constraints, these annotations

The plane was ***diverting*** *around weather formations over the Java Sea when contact with air traffic control (ATC) in Jakarta was* ***lost****.*

wh	aux	subj	verb	obj	prep	obj2	?	Answer
What	was		being diverted		around		?	*weather formations*
What	was		diverting				?	*The plane*
What	was		being diverted				?	*The plane*
What	was		lost				?	*contact with air traffic control*
Where	was	something	lost				?	*over the Java Sea*

Table 1: QA-SRL question-answer pairs from the development set of the QA-SRL Bank 2.0 (FitzGerald et al., 2018). We constrained the questions with a non-deterministic finite automaton (NFA) encoding English clause structure for question autocomplete and auto-suggest. This facilitated high-quality, high-coverage annotation at scale while providing the expressiveness to represent the semantic role relations within each sentence.

are somewhere between our work on human-in-the-loop parsing and question-answer meaning representations. The constrained task and high coverage allow us to train high-quality QA-SRL predictors and enables future work on semantic role induction (Section 5.1) and controlled question generation (Section 5.2).

Takeaways Our results over the course of these projects suggests that we should search for tasks in a “goldilocks zone”: Their scope should not be so constrained or beholden to prior theory as to be unintuitive, but not so unconstrained that it is hard to get exhaustive and reliable annotation of interesting phenomena. As annotation constraints depend on *some* prior theory of the phenomena to be captured, these constraints need to be carefully chosen so as to minimize arbitrary assumptions in the task setup and make sure the task is natural for annotators. In the case of QA-SRL, the prior theory we incorporated is a small grammar fragment of English encompassing QA-SRL questions. Our findings support that QA-SRL, with the annotation aids developed in FitzGerald et al. (2018), strikes a good balance of the Four Principles.

5 Theory: From Language, Structure

In this section, I will describe two projects which show how QA-SRL can be used to build a data-driven theory which is directly applicable in downstream tasks.

5.1 Inducing Semantic Roles Without Syntax

Michael and Zettlemoyer (2021) show how to use QA-SRL to automatically induce an ontology of semantic roles, leveraging a key insight: the *set* of QA-SRL questions that are correctly answered by a given answer span identifies an underlying semantic role through its syntactic alternations, which are representative of the phenomena that a semantic role ontology like PropBank is designed to explain. We leverage this insight by using a trained QA-SRL question generator as a data simulator, generating a full distribution over (simplified) QA-SRL questions for each argument of a verb appearing through an entire corpus. Clustering these distributions of questions according to a simple maximum-likelihood objective yields a set of discrete semantic roles that exhibits high agreement with existing resources (see Table 2). This presents an approach which could potentially be used to develop semantic role ontologies in new domains where they are not currently available, with directions for improving QA-SRL data toward the end of automatically inducing better semantic roles.

Labels	Questions	
A1 (98%)	What is given?	.30
	What does something give something?	.21
	What does something give?	.20
	What is something given?	.11
A0 (98%)	What gives something?	.44
	What gives something something?	.27
	What gives something to something?	.08
A2 (94%)	What is given something?	.28
	What does something give something to?	.18
	What does something give something?	.14
	What is given?	.09
	What is something given to?	.07
TMP (46%),	When does something give something?	.20
ADV (22%),	How does something give something?	.09
MNR (12%)	When is something given?	.09
	When is something given something?	.09
PNC (30%),	Why does something give something?	.18
ADV (22%),	Why does something give up something?	.07
TMP (14%)	Why is something given something?	.07

Table 2: Roles for *give* produced by Michael and Zettlemoyer (2021). For each predicate, we cluster its arguments in PropBank based on the similarity of the distributions of QA-SRL questions our model generates. In this case, core arguments are captured almost perfectly, exhibiting both passive and dative alternations.

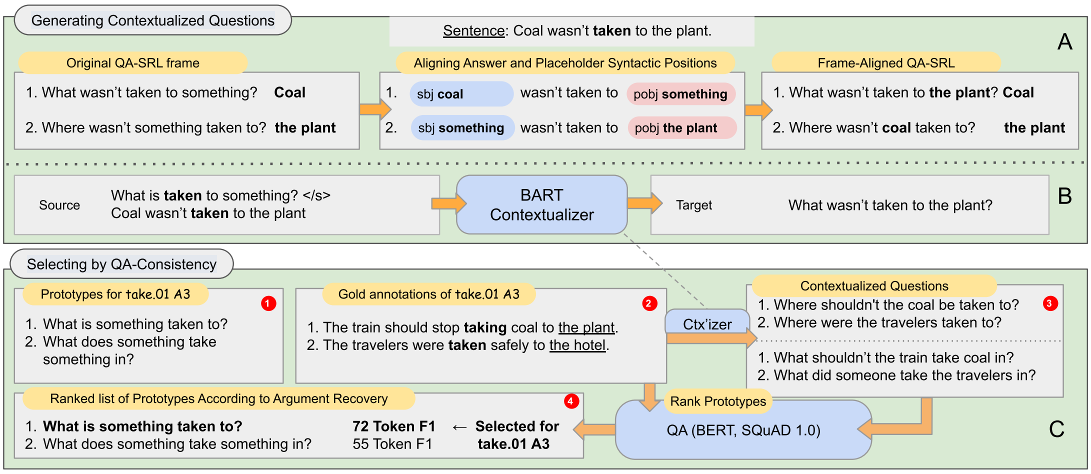

Generating Contextualized Questions
Sentence: Coal wasn't taken to the plant.
A
Original QA-SRL frame
1. What wasn't taken to something? Coal
2. Where wasn't something taken to? the plant
Aligning Answer and Placeholder Syntactic Positions
1. sbj coal wasn't taken to pobj something
2. sbj something wasn't taken to pobj the plant
Frame-Aligned QA-SRL
1. What wasn't taken to the plant? Coal
2. Where wasn't coal taken to? the plant
Source
What is taken to something? </s>
Coal wasn't taken to the plant
BART
Contextualizer
Target
What wasn't taken to the plant?
B
Selecting by QA-Consistency
Prototypes for take.01 A3
1
1. What is something taken to?
2. What does something take something in?
Gold annotations of take.01 A3
2
1. The train should stop taking coal to the plant.
2. The travelers were taken safely to the hotel.
Ctx'izer
Contextualized Questions
3
1. Where shouldn't the coal be taken to?
2. Where were the travelers taken to?
1. What shouldn't the train take coal in?
2. What did someone take the travelers in?
Ranked list of Prototypes According to Argument Recovery
4
1. What is something taken to? 72 Token F1 ← Selected for take.01 A3
2. What does something take something in? 55 Token F1
Rank Prototypes
QA (BERT, SQuAD 1.0)
C

Figure 3: Overview of Pyatkin et al. (2021)'s approach. The natural correspondence between QA-SRL questions and semantic roles allows us to use QA-SRL question templates in a planning step to successfully generate questions for any PropBank semantic role, even when the corresponding argument doesn't appear in the source sentence (a situation never encountered in training data). **A: Construction of Frame-Aligned QA-SRL** using syntactic information inferred by the autocomplete NFA from FitzGerald et al. (2018), *i.e.*, leveraging our (minimal) theoretical assumptions about argument structure. **B: Contextualizing questions** by feeding a prototype question and context into a neural model that outputs a Frame-Aligned QA-SRL question. **C: Selecting prototype questions** by testing each prototype (1) against a sample of arguments for each role (2). After contextualization (3), each question is fed into a QA model and we choose the prototype that most often recovers the correct argument (4).

5.2 Asking it All: Generating Contextualized Questions for any Semantic Role

Pyatkin et al. (2021) use QA-SRL to build a controllable question generation system. The task is to generate fluent questions asking about the arguments corresponding to specific semantic roles in context (see Figure 3 for an overview). The challenge is a lack of training data, as QA-SRL questions are not fully natural and are not annotated for roles which aren't expressed in a sentence. We leverage two key insights: First, we find that QA-SRL questions generally correspond to the same role across many contexts. So we prime our question generation system with a template QA-SRL question corresponding to the correct role, leading it to generate semantically correct questions even when the answer isn't present in the sentence. Second, we use the syntactic structure of QA-SRL questions to align the placeholders (*someone*, *something*) in each question with the answers of other questions, translating QA-SRL questions into more fluent ones closer to those in QAMR.

Takeaways Together this work illustrates not only the promise for the development of large-scale ontologies in a data-driven way (Section 5.1), but it also illustrates how having these ontologies computationally grounded in the phenomena they are designed to explain, *i.e.*, question-answer pairs, facilitates ontology's the downstream use (Section 5.2). It's not hard to imagine next steps incorporating an induced ontology of semantic roles into Pyatkin et al. (2021)'s system to obviate the need for a pre-specified role ontology altogether.

6 Concluding Thoughts

I have proposed *scalable, data-driven theory* as a Pragmatist paradigm for scientific progress in NLP. To develop scalable theories, one should:

1. Collect carefully-scoped data that directly represents a phenomenon of interest while imposing minimal prior theoretical assumptions,
2. Increase the data's scale and coverage using a learned black-box data simulator,
3. Induce comprehensible models of this high-coverage data with machine learning, and
4. Examine the results to debug and improve the theory and data, progressing our scientific understanding of the phenomenon of interest.

Using QA-SRL, I have shown how to leverage black-box data simulation together with simple probabilistic modeling to automatically induce an ontology of semantic roles which is directly and comprehensibly grounded in phenomena that the theory of semantic roles is meant to explain. This not only lays the groundwork for new scalable theoretical developments in semantic representation, but can serve as an example to guide future work on scalable theories in other domains.

Why now?

The justification for building scalable, data-driven theories can be summarized as follows:

1. To build systems which generalize in controllable, predictable ways, we need comprehensible theories of their desired behavior.
2. However, the behaviors we wish to produce in AI and NLP are too complex for us to easily write down theories of how they should work.
3. So instead, we must use machines (*i.e.*, statistical models) to construct our theories on the basis of data in a scalable way. The role for the scientist here is twofold:
   - to carefully determine the scope of the phenomena to be explained and curate the data accordingly, and
   - to define the meta-theory which relates the learned theory to the data.

This argument could have been made at any point in the history of NLP, so why do I make it now?[4] I think the argument would have been viewed as premature in the *era of underfitting* prior to the deep learning revolution. Statistical models like CRFs (Lafferty et al., 2001) struggle even in-distribution on tasks like syntactic and semantic parsing, let alone complex end tasks involving question answering or language generation. The problem at that time was to build models expressive enough to perform well while tractable enough to learn from data. Pre-neural systems were weak enough that many thought they would benefit from hand-curated linguistic resources like PropBank (Palmer et al., 2005).

With deep learning, these factors all changed: the limits of hand-curated resources like PropBank have been surpassed, and neural models fit all kinds of data distributions, leaving us face-to-face with the problem of generalization and the need for data-driven theory. Furthermore, we have new tools for data simulation; the role induction algorithm in Michael and Zettlemoyer (2021) would not have been workable without a neural model to simulate dense annotation of QA-SRL questions. So we are finally in a position to make such theories scalable.

Looking forward

As argued above, a critical role for the scientist in developing data-driven theories is to define scopes of phenomena to be explained, carving linguistic behavior at useful joints. I hope to have demonstrated that the concept of *semantic roles* provides such a useful scope, where its corresponding phenomena (as QA-SRL) can be effectively annotated at scale (Section 4.3), tractably modeled with a comprehensible theory (Section 5.1), and used for downstream tasks (Section 5.2). Moving forward requires carefully choosing more such useful concepts and using them to scope phenomena, define and induce theories, and tie these data and theories into downstream applications.

Extending the paradigm of scalable theory to more facilities of language (*e.g.*, syntax, word sense, or coreference) and more complex phenomena (*e.g.*, representations of world knowledge, common sense, or reasoning) remains a major challenge. As the scope of the phenomena to be represented increases, greater annotation constraints will be necessary in order to ensure that these phenomena are adequately covered. However, doing so while maintaining theoretical minimalism is challenging. My hope is that scalable theories of narrowly-scoped subphenomena (*e.g.*, semantic roles) will provide constraints that make more complex tasks tractable to exhaustively annotate, without introducing the same problems as in the Rationalist paradigm where inconsistencies, underspecification, and arbitrary theoretical choices limit the usefulness of the data. In this way, it may be possible to bootstrap from narrowly-scoped theories into progressively broad accounts of language structure, meaning, and intelligent behavior.

At this point, such talk is speculation. It is unclear how data-driven theory will generalize to more complex tasks. However, in this work I hope to have provided an argument this kind of work is at least worth attempting, and perhaps laid some groundwork and principles which can be used as a starting point for it to be done in the future.

[4] Similar arguments have been made before in grammar engineering (Oepen et al., 2004; Flickinger et al., 2017) and the Decompositional Semantics Initiative (White et al., 2016), while in linguistic typology, Haspelmath (2010)'s *framework-free grammatical theory* makes similar points about the relationship between data and theory. My approach differs from these in my focus on applications in NLP where the vastness and complexity of the domain becomes more of a challenge.

Acknowledgments

Thanks to my PhD thesis advisor Luke Zettlemoyer, as well as reading committee members Noah A. Smith and Emily M. Bender, and committee member Shane Steinert-Threlkeld. Many thanks also to my collaborators on the projects reviewed in this note, including Ido Dagan, Luheng He, Gabriel Stanovsky, Valentina Pyatkin, Paul Roit, and Nicholas FitzGerald, and others who have done essential QA-Sem work following on QA-SRL, including Ayal Klein and Daniela Weiss, as well as the many annotators who have contributed to building these datasets. Thanks also to my brother Jonathan Michael for introducing me to Pragmatism and Ari Holtzman for helpful and engaging discussions about it. Finally, thanks to the anonymous reviewers for helpful comments on what I should include in this note to round out the discussion. See Michael (2023) for more detailed acknowledgments for my thesis work.

References

Hannah Youngeun An and Aaron Steven White. 2020. The lexical and grammatical sources of neg-raising inferences. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 386–399, New York, New York. Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity.

Emily M. Bender and Guy Emerson. 2021. Computational linguistics and grammar engineering. In Stefan Müller, Anne Abeillé, Robert D. Borsley, and Jean-Pierre Koenig, editors, *Head-Driven Phrase Structure Grammar: The handbook*, Empirically Oriented Theoretical Morphology and Syntax, pages 1101–1148. Language Science Press., Berlin.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *COLING-02: Grammar Engineering and Evaluation*.

Jonathan Berant, Noga Alon, Ido Dagan, and Jacob Goldberger. 2015. Efficient global learning of entailment graphs. *Computational Linguistics*, 41(2):221–263.

George E. P. Box. 1976. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending:strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics.

David Carter. 1997. The TreeBanker: a tool for supervised training of parsed corpora. In *Computational Environments for Grammar Development and Linguistic Engineering*.

Kenneth Church. 2007. A pendulum swung too far. *Linguistic Issues in Language Technology*, 2.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Edsger W. Dijkstra. 1974. Ewd 447: On the role of scientific thought. In *Selected Writings on Computing: A Personal Perspective*, pages 60–66. Springer-Verlag. Book published in 1982.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Dan Flickinger, Stephan Oepen, and Emily M. Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. In *Handbook of Linguistic Annotation*, pages 353–377, Dordrecht. Springer Netherlands.

William Gantt, Lelia Glass, and Aaron Steven White. 2021. Decomposing and recomposing event structure. *CoRR*, abs/2103.10387.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019a. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019b. Question answering is a format; when is it useful? *CoRR*, abs/1909.11291.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Haspelmath. 2010. Framework-free grammatical theory. In *The Oxford Handbook of Linguistic Analysis*, Oxford, UK. Oxford University Press.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Luheng He, Julian Michael, Mike Lewis, and Luke Zettlemoyer. 2016. Human-in-the-loop parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2337–2342, Austin, Texas. Association for Computational Linguistics.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

William James. 1907. *Pragmatism: a New Name for some Old Ways of Thinking*. Project Gutenberg.

Dan Jurafsky and James Martin. 2008. *Speech and Language Processing*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Douglas B. Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):32–38.

Vladimir Lifschitz. 2008. What is answer set programming? In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1594–1597. AAAI Press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

John McCarthy. 1976. An example for natural language understanding and the ai problems it raises.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Mark Johnson, and Mark Steedman. 2023. Smoothing entailment graphs with language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 551–563, Nusa Dua, Bali. Association for Computational Linguistics.

Julian Michael. 2023. *Building Blocks for Data-Driven Theories of Language Understanding*. Ph.D. thesis, University of Washington.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Julian Michael and Luke Zettlemoyer. 2021. Inducing semantic roles without syntax. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4427–4442, Online. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2023. The alignment problem from a deep learning perspective.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D. Manning. 2004. LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 2:575–596.

OpenAI. 2023. Gpt-4 technical report.

Martha Palmer, H. Dang, and C. Fellbaum. 2006. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13:137–163.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Stefan Riezler. 2014. Last words: On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, 40(1):235–245.

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, and Rui Wang. 2011. The ACL Anthology searchbench. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 7–13, Portland, Oregon. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Rich Sutton. 2019. The bitter lesson.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Aaron Steven White. 2021. On believing and hoping whether. *Semantics and Pragmatics*, 14(6):1–21.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of s-selection. *Semantics and Linguistic Theory*, 26:641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, pages 221–234, Amherst, MA. GLSA Publications.

Aaron Steven White, Kyle Rawlins, and Benjamin Van Durme. 2017. The semantic proto-role linking model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 92–98, Valencia, Spain. Association for Computational Linguistics.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.