I am broadly interested in the computational foundations of intelligent behavior through the lens of natural language. The overarching theme of my research is centered around developing algorithms and theories that make natural language processing (NLP) systems more general and generalizable, i.e., enabling them to adapt and handle a broader space of challenges or situations. Humans have seamless generalizability – one moment we’re playing chess, the next moment we are responding to an emergent situation. Within AI, however, our successes have been mostly limited to narrowly-defined tasks (narrow vs extreme generalization; Fig.1). While AlphaGo, for example, has aced the game of Go [1], it is unable to solve other related problems (such as, explaining the moves that it makes or solving another similar puzzle). The progress toward the ambitious goal of generalizable models requires rethinking different stages of the AI pipeline. In particular, during my past research, I have pursued this vision on three complementary axes:

- (A) Inducing generality in task formulations by defining and tackling a broader scope of tasks and abilities and enabling us to measure more realistic senses of generalization [3; 4].
- (B) Enriching representations that support model generalization by utilizing cheap signals available in the wild, independent of any downstream task [5; 6].
- (C) Arming models with general-purpose reasoning paradigms to enable them to infer new findings and communicate their unknowns, in a way that supports a broad-ranging spectrum of tasks [7; 8; 9].

Needless to say, these three aspects of AI design are not disjoint, but rather inter-dependent. While each section focuses on a particular angle, the presented works belong to more than one camp.

A Defining and Tackling Tasks with Broader Definitions

How can one formulate setups that measure a broader generality (Fig.1)? One instance of this is our recent approach to solving Question Answering (QA) [3], a popular setup to assess computers’ ability to understand language and reason with language [10]. Over the years, our community has produced many datasets with a variety of formats (multiple-choice, reading comprehension, and so on; Fig.2). The existence of different variants of QA has, sadly, has resulted in research silos: most of the past works focus on one QA dataset [11; 12], or at best, several datasets of the same format [13; 14]. In our EMNLP’20 work [3], we argued that such boundaries between QA formats are unnecessary by empirically showing that there is indeed transfer across seemingly distinct QA variants, i.e., supervision with one format helps QA systems perform on questions in another format (different sub-trees in Fig.2). Intuitively, the abilities needed to answer questions are not bound to task or dataset formats. Building on top of key intuition, we proposed a single format-agnostic QA model, UNIFIEDQA, that performs well across 20 QA datasets spanning 4 distinct formats. This system compromised little compared to format-specific models while showing remarkable generalization to other unseen datasets. At the time of its publication, UNIFIEDQA was the most general QA model which achieved new state-of-the-art performance on 10 different NLP benchmarks.

The success of UNIFIEDQA has had several ripples of impact. First, it alleviated the conceptual barriers for building unified models. In less than two years, there have been several important follow-ups that extend our core idea to different problem spaces [15; 16; 17; 18; 19; 20]. Second, the empirical success of UNIFIEDQA is reproduced on tasks and datasets that did not exist at the time [21; 22; 23; 24] – strengthening our earlier intuitions on the generality of our approach. In one particular case, UNIFIEDQA was shown to have stronger generalization than GPT3, a model that is 16× larger than UNIFIEDQA [25].

Despite the success of models like UNIFIEDQA they fail to generalize outside the space of QA problems. In a recent work [4], we introduce a formulation for studying task-level generalization (i.e., generalization to unseen tasks).
In our proposed setup, each task consists of an instruction document that defines how an input text is mapped to an output (Fig.3). A hypothetical model equipped with understanding (and executing) language instructions should be able to generalize to any task that can be defined in terms of natural language (Fig.3). This is a broad formulation that subsumes many tasks in NLP. To study this setup we built Natural-Instructions, a dataset of language instructions for over 1k tasks. We use this dataset to benchmark cross-task generalization (i.e., train models on a subset of the tasks and evaluate them on the remaining ones), across a diverse range of tasks—a setup that was not possible previously. Our experimental results verify the value of instructions in producing generalization: models that use task instructions gain increasingly better generalization when they get to observe more tasks. In particular, a small model obtains a level of generalization to unseen tasks that is on-par with GPT3 despite being 1200× smaller.

B Generality Enabled by the Representation of Task-Independent “Incidental” Signals

Knowledge representation—either in symbolic or distributed forms—is the foundation on which AI systems function. The richer and broader this foundation is, the more general the model will be. The revolution of the past few years is about building effective representations by large-scale pre-training of models [26; 27] on “incidental” signals—collections of signals that exist in the wild independent of downstream tasks [28], such as freely available text on the web. Despite the success of pre-training, there are many aspects of language that are not effectively covered by them. That is why part of my focus has been about innovating approaches to exploit untapped incidental signals.

Understanding of time is a crucial element of natural language and many downstream applications [29; 30] which is not addressed by the existing approaches. Consider the example of Fig.4. Humans know that a typical vacation is likely to last at least a few days, and they would choose “will not” to fill in the blank for the first sentence; however, for the second sentence which contains a slight change of context (“vacation” → “walk outside”) people typically prefer “will”. Recent pre-trained models cannot handle such examples, partly because of reporting bias (people rarely mention unnecessary details, such as the duration of “brushing teeth”). In our ACL’20 paper [6], we augmented conventional pre-training with a temporal objective that incorporated knowledge of temporal events. This objective fuses symbolic world knowledge about time (the relation between temporal units and events) with the distributional statistics mined from free-form text. For instance, it forces the ordinal relation of temporal units (such as, “seconds” < “minutes” < “hours”). It also induces interdependence between the inferred temporal dimensions (temporal duration, frequency, typical time, and so on). For example, “I brush my teeth every morning” which indicates the frequency of the “brushing teeth” event, implies an upperbound for the duration of the same event. By incorporating this intuition into the pre-training objective, we built a language model that is informed of the temporal properties of events (Fig.5). Our construction is independent of any target task, as verified by its generalization to several extrinsic benchmarks that require an understanding of time. This was the first work to augment language models pre-training with a temporal objective that relates various temporal units and events.

Understanding entities and abstracting over them is another important ability that appears in applications [31]. Consider the sentence that contains “Bloomberg” (Fig.6). How should a model know the semantic type(s) of “Bloomberg”? This is non-trivial since this particular mention can be a politician, businessman, magazine, company, and so on, depending on the context it appears in. The traditional approach to solve this involves supervising models with datasets that have expert annotations for entity types according to a fixed and often limited taxonomy of types [32; 33]. Models built according to this paradigm are constrained to the type taxonomy they were supervised with. For example, a model trained to recognize persons, cannot easily be adapted to distinguish, say, politicians and entrepreneurs.

To address this limitation, in our EMNLP’18 paper [5] we build a model for typing entities without relying on any expert-annotated labeled datasets. We use readily-available resources such as Wikipedia that cover many signals needed to disambiguate entity mentions. Wikipedia is a rich resource that contains millions of entities and a
link structure that reveals their types. How can we infer the semantic type(s) of an entity mention, even if it is not in Wikipedia? Our approach involves forming “definitions” for each type and subsequently, checking if the mention aligns with the definition of a given type. In particular, each semantic type is “defined” by its instances and the context in which they are used (e.g., politician is partially defined by all the sentences in which Elizabeth Warren is described). Given this representation of entity types, we cast the semantic typing problem as a nearest-neighbor algorithm that uses contextual similarities between a given mention and Wikipedia entities of known types (Fig. 6). The result is Zoe (zero-shot entity-typer), a system that uses no manually-labeled data as supervision. Since this construction was not reliant on any particular dataset as a source of supervision, Zoe generalizes to several popular benchmarks on which it was shown to perform competitively with state-of-the-art supervised systems that are restricted to the taxonomies they were supervised with.

C Generalization via Reasoning-Driven Design

Reasoning1 has long been believed to be a source of generalization in human judgments about new problems and environments. Since the dawn of AI, this intuition has motivated frameworks with relatively general and abstract primitives for decision-making [35; 36]. With the rise of statistical machine-learning, many of the ideas in the so-called “symbolic” AI camp are forgotten, even though they possess positive attributes (e.g., ease of interpretability) that are non-trivial in the recent Deep-Learning-based technologies.

Part of my research has been about marrying state-of-the-art technology models with the appealing properties of classical AI in a way that leads to a new level of generality [7; 37]. For example, our AAAI’18 paper [7] introduces a model that casts question answering as a subgraph search problem over semantic representations extracted from statistical models, such as semantic role and coreference annotators [38; 39]. Enabled by the task-independence of this underlying representation and the reasoning on top of it (sub-graph search), our system showed notable generalization across several QA benchmarks compared to black-box state-of-art systems at the time. These systems were effect components of Aristo [40], a larger QA system developed at Allen Institute for tackling elementary-school science exams. Since then, these ideas have inspired much follow-up work on multi-hop reasoning based on similar paradigms or modernized learning architectures [41; 42].

A realization of reasoning that I have explored is interactive communication for the sake of making conclusions (cf. Footnote 1). We, humans, often solve complex tasks by breaking them down into manageable sub-tasks, solving them in interacting—in natural language— with other people or automated agents whose respective skill-sets we are familiar with. Can AI systems learn to do the same? In our NAAACL’21 work [8], we introduced a general-purpose framework that casts complex tasks as textual interaction between existing, simpler QA modules (Fig. 7). Based on this conceptual framework we proposed ModularQA, a system that can perform multi-hop and discrete numeric reasoning. ModularQA was the first modular system that worked on several notable benchmarks at the time and achieved and on par with other dataset-specific modular systems. The ability to (learn to) interact with existing systems leads to a model that is more versatile and explainable than state-of-the-art black-box (uninterpretable) systems, at the cost of a little overall accuracy.

Since reasoning remains an open-ended aspiration, one has to hunt down aspects of it that are not captured by the existing evaluation benchmarks (i.e., the need to discover and characterize the untouched branches of the task hierarchy; Fig. 8). In my prior work, I built several reasoning-driven benchmarks [43; 44; 45] via careful crowdsourcing designs, some of which are widely used and have become part of popular leaderboards.2 As a recent example, we introduced a QA dataset with implicit decompositions in our TACL’21 paper [9]. In most of the prior multi-hop

1 While “reasoning” has been studied for over a millennium, its nature remains a subject of debate among cognitive and social scientists. Until a few decades ago reasoning was considered a means to think better on one’s own (such as making deductive conclusions). The recent theories suggest that reasoning is often done in (and evolved through) interaction with others. A recent work defines it as “an act of producing arguments for explaining (justifying) oneself or convincing others” [34].

2For example, MultiRC [45] is part of SuperGLUE: https://super.gluebenchmark.com
QA datasets [43; 46], the language of questions explicitly describes the process for deriving the answer. Take the example of “Was Aristotle alive when the laptops were invented?” which explicitly specifies the required reasoning steps. However, in many real-life questions, the necessary reasoning is implicit. For example, the question “Did Aristotle use a laptop?” (Fig.8) can be answered using the same steps, but the model must make several implicit inferences. For example, a system needs to infer that “X using Y” necessitates co-existence of X and Y at the same time. Answering implicit questions poses several challenges compared to answering their explicit counterparts. My hope is that posing such challenges will motivate the development of more general models that rely less on the surface cues of input questions.

**Future Work**

While the revolution of the past decade was mainly about “representation” provided by the pre-training language models [26; 27], I hypothesize that this decade will be about the generality of our models. Speculating about our long-term progress, by the end of the next decade all the isolated application of AI today (Alexa, search engines, movie and product recommender systems, self-driving software, etc.) will become part of a broad and homogenized AI system. The futuristic personal assistants that will seamlessly integrate any task that we currently accomplish via distinct applications (emails, calendar, weather, maps, etc.) and devices (phone, laptop, etc.) The current AI is far from this long-term milestone as they are limited to narrow scopes of problems. To move toward more broad-ranging AI systems, in the near-term I plan to tackle the following fronts:

**Toward Broader Formulations and Understanding of Generalization**

In near term, I plan to build upon the setups in §A and develop broader problem frameworks that support aspects such as multi-linguality and multi-modality (performing visual or voice tasks that can be described in language). Such developments will enable any person to conveniently communicate with audio and video editors via language commands – just describe your desired effect and the software will do it!

The future progress on generalization necessitates a holistic understanding of its scaling laws as a function of various parameters: what tasks (don’t) benefit from one another, whether models generalize to other problem domains such as embodied environments in the robotics community, whether they generalize along the abstractness axis (e.g., applying abstract and high-level ethical principles to specific scenarios [47]), the role of pre-training our models, and so on. Investigating such questions will guide future steps toward this challenging setup.

**Richer Representation Enabling More General Models**

The progress of the past few years enabled by large-scale pre-training [26; 27], I hypothesize, will continue to yield better representations. There are so much untapped incidental signals in the wild that we haven’t factored in yet. For example the implicit interactions on the web: Opinion columns rebutting each other, and science articles building upon earlier works, etc. Similarly, there are untapped signals for learning to decompose complex problems (§C): many computer codes (say, on Github) divide problems into existing sub-functions, math papers prove theorems by reducing them into known lemmas, and so on. Beyond the Web, another untapped frontier of cheap data is the environment around us: how we navigate our physical world and interact with others. Harnessing such incidental signals will further strengthen the foundation of future models.

**Informed and Communicated “Ignorance”**

For our models to generalize to off the beaten path (environments with many unknowns) and discover new findings, the models need to recognize their ignorance (know what they don’t know). This is a necessity for having reliable text-based agents that don’t make up hallucinated statements. Oddly, in classic AI that represented the world with symbols, this property was given. But in the context of recent technologies enabled by neural networks, this is non-trivial and has received too little attention, only for narrow tasks setups. To bring more attention to this problem, we need formulations of the problem that capture a broad range of tasks and abilities. A successful attempt to solve this problem will likely require a novel marriage of modern NLP with inspirations from the symbolic AI literature.

Furthermore, a model that is informed about its “ignorance” needs to articulate the unknowns in a language that is understandable to the world (human users or other AI systems). An instance of this was show-cased as MODULARQA (§C) which relied on assumptions about its target tasks and domains. There is plenty of room for progress in generalizing this idea into a unified interactive inquiry mechanism for a wide range of unknowns and domains.
References


