Research Statement

Daniel Khashabi

December 2019

My research focuses on the computational foundations of *intelligent behavior*, through the lens of *natural language*. I am interested in the development of *theories*, *algorithms* and *systems*, through unified methodologies. My goal, in the long run, is to develop capabilities for *natural language understanding* (NLU), i.e., enabling computers to understand language, as close as possible to the ways in which humans would interpret it. In the near-term, I am motivated by applications of language processing, empowered by machinery like *deep learning*. The path to this goal covers a variety of subfields: from foundational questions in *machine learning*, *knowledge-representation* and *reasoning* to experimental paradigms and large-scale system development.

There are two key subjects of my research: *humans* and *machines*. While we often work to improve *machines*, the ultimate goal is to help *humans*. Posed as two broad questions that motivate my research:

1 (§A) *From Humans to Machines:* Can we build *systems* that understand and *reason* with *human* language?

The answer to this question can be explored at various stages: (i) defining a set of tasks that illustrate the expected language understanding capabilities is a natural first step as they help identify the open challenges, provide means to measure progress and encourage tackling the blind spots. For example, the task of answering questions, such as the one posed in Fig.1. (ii) Systems designed to answer such questions (§A.2) should deal with many challenges: Human language is quite complex, since it is both *ambiguous* and *variable* (Fig.2); Additionally, small changes in language can make significant differences in its meaning (e.g., changing "longest" to "shortest" in Fig.1). A system working with natural language should be robust to all such complexities in language. (iii) Beyond empirical work, to support our intuitions we also need theories and formalism (§A.3) that explain the extent and limitations of our empirical understanding.

Figure 1: An example language question: a system answering such questions should be robust to its variations.

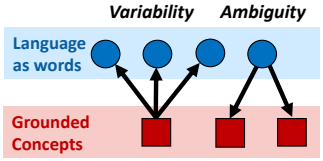

Figure 2: Left: a single concept described by different words (variability). Right: a single word referring to multiple concepts (ambiguity).

2 (§B) *From Machines to Humans:* Can we use *machines* equipped with NLU for the common good?

There is plenty of room for technologies like NLU to help people. For instance, one major concerning trend in the age of information is *echo-chambers*. Many of us are trapped in bubbles of like-minded groups: we only interact with and hear those we tend to share similar views with (5). Unfortunately there are grave consequences to this: the breakdown of public discourse on national-level issues that can lead to ideological divide among citizens and weakening of democracy. However, we might be able to help individuals see a wide variety of *perspectives* (§B.1) and alleviate *echo-chambers* by inventions based on language technologies.

The following sections represent vignettes from my past and ongoing explorations in response to the questions above. These are results of contributions published in major AI/NLP conferences.

A Reasoning about (and with) Natural Language

A.1 Defining effective (and useful) problems. Defining language tasks that both have broad coverage and lead to meaningful measure of the progress is not an easy exercise. First, any dataset carries only a limited

linguistic complexity. Additionally, it has been recently shown that many benchmark datasets have malicious biases that can be exploited by algorithms (12) leading to (misleadingly) high performance.

Part of my work is devoted to defining datasets that capture aspects of language understanding that are still challenging for state-of-the-art systems (7; 16). For instance, building datasets that require *temporal commonsense understanding* (16) — the implicit inference about temporal properties (duration, frequency, etc) of event mentions in natural language (see Fig.3). The best existing models are still far behind human performance on this task (a gap of $\approx$ 30%)[1]. Such gaps pose a challenge (and an opportunity) for the NLP community to address.

Consider the following two events with a similar surface:

1. *"taking a vacation"*
2. *"taking a break outside"*

The former could takes weeks, while the latter could take few minutes.

Figure 3: An example of "temporal commonsense".

[1] https://leaderboard.allenai.org/mctaco

A.2 Building reasoning systems. The Aristo project at the Allen Institute for AI, a project I have closely collaborated with in the past several years, aims to build a machine that can understand elementary-school science (3). To measure progress, the project uses standardized tests as its challenge since it is a measurable target and such questions often require a wide variety of nontrivial reasoning abilities.

A major challenge involves what is called "multi-hop" reasoning — the ability to chain pieces of information in order to draw a conclusion. For example, to answer the question in Fig.1 a machine solver needs to find multiple semi-disjoint bits of information: an understanding of NYC's location, orbital events and their connection to day/night durations. Our progress resulted in multiple generations of reasoning systems that cast question answering as a subgraph search problem over some semi-structured representations, e.g., database tables (see Fig.4) (6) or more complex semantic representations (8). Since our earlier work, these ideas have inspired much follow-up work on multi-hop reasoning based on similar paradigms (13) or modernized learning architectures (15).

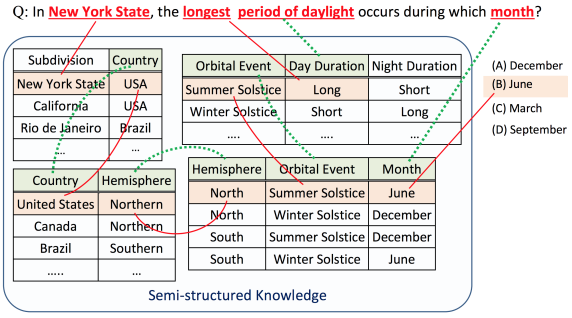

Figure 4: Depiction of reasoning process done by our system, for the example provided in Fig.1. The system searches for the best support graph (chains of reasoning) connecting the question to an answer, in this case *June*.

A.3 Formalizing limitations of multi-step reasoning. In addition to empirical progress, we need theoretical work that can support (and explain) empirical findings. In recent work (9) we present the first formalism that addresses the empirical intuition that the accuracy of multi-hop systems significantly drop even with small increase in the number of required reasoning steps. Our framework allows one to *quantify the effect* of linguistic imperfections (e.g., ambiguity, redundancy). We formally define a noisy (parametric) channel between the space of "concepts" and the space of "words". By applying this framework to a special class of simplified multi-hop decision problems we derive rigorous intuitions and impossibility results. For instance, if a query requires a moderately large (logarithmic) number of hops, no reasoning system operating over a noisy graph (of linguistic knowledge) is likely to succeed. This highlights a fundamental barrier for a

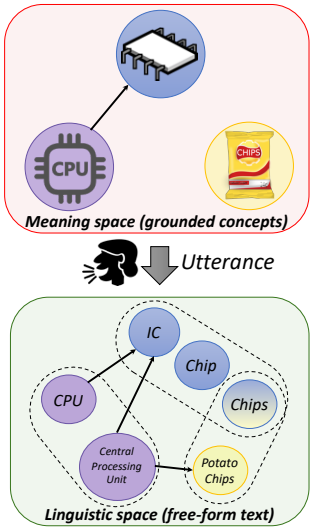

Figure 5: The noisy channel between between meanings (concepts) and linguistic space (words) capturing linguistic imperfections: each meaning (top) can be uttered in many ways as words (bottom), and the same word can have multiple meanings.

class of reasoning systems. We expect our findings to have important implications on how we study problems in language comprehension.

B Natural Language Systems Assisting Human Decision-Making

One of the biggest challenges facing us in the age of data is *information pollution*: "the contamination of information supply with irrelevant, redundant, unsolicited, hampering and low-value information" (14). Societies are increasing in diversity, bringing together people with different backgrounds and perspectives about contentious questions. On the other hand, we are exposed to an increasing amount of content (news, social media, ads, etc) in our daily lives, and a significant portion of them distort the reality (4) (intentionally or unintentionally) by showing only part of the picture. This is further exacerbated by personalized social media algorithms which tend to deliver content that corroborate our existing beliefs and encourage the formation of "bubbles" of like-minded people. This motivates several of my past & ongoing projects.

B.1 Helping individuals see issues from different perspectives. In a recent project, we are building systems that encourage an inclusive and holistic view of many challenging issues. Take the following controversial question: *"can animals have lawful rights?"* There might not be a simple answer to this question, as there are many different aspects to address.

Our goal in this project is to explore better ways to organize ideas and perspectives. We would like to help individuals see the other sides of the aisle, by organizing different aspects of the problems.

To start with, we characterize the core NLP challenges required to solve the perspective discovery problem. To consolidate our task formulation and facilitate research in this direction, we construct a dataset of claims, perspectives and evidence documents (2). In the task, a system is expected to discover all the relevant perspectives (supporting or undermining), followed by extracting all the pieces of evidence that substantiate each perspective (Fig.6). Our dataset has already gained attention within the community, as several teams are building upon it.

To make the idea more accessible, we developed a platform[1] that simulates this end-to-end process of minimal perspective discovery (1). We are actively studying ways to make this platform reliable and accessible anyone who might benefit from it.

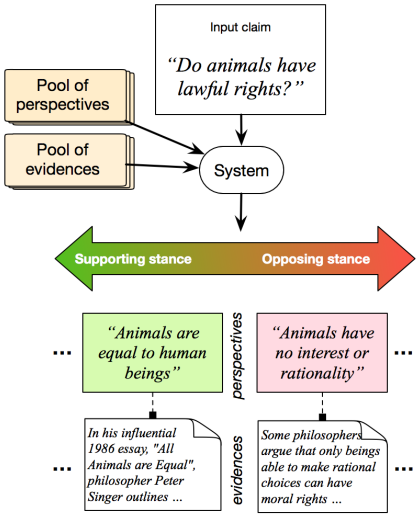

Figure 6: Given a *claim*, the system is expected to discover different *perspectives* that are substantiated with *evidence* and their *stance* with respect to the claim.

[1] http://perspectroscope.com

C Future Research Directions

In future research, I plan to build on the following research directions:

C.1 Better methodologies for measuring the progress. While past few years the NLP community has made significant progress in better representational techniques and model design (§A), our evaluation paradigms are lagging behind. We are increasingly aware of the biases and artifacts in machine-learning pipelines that potentially distort the quality of evaluation (10; 12). There is a need for approaches that better quantify whether the models have learned the intended abilities rather than over-fitting potential idiosyncrasies of

a dataset. While several of my ongoing works address this angle, for example by measuring robustness to out-of-distribution perturbations (Fig.1). That being said, there are many open problems with respect to the generalization of systems.

C.2 Explainablity via decomposed decisions. One of the underexplored components in the modern AI technologies (especially *deep learning*) is the lack of high-level understanding and explainability. Suppose a state-of-the-art model makes mistake in the question of Fig.1. How do we know what is going wrong? the model doesn't understand the question? or it doesn't know where NYC is? One high-level approach is to design models that could *decompose* their decision into into smaller (ideally, interpretable) steps. For instance, in the question in Fig.1, rather than a single-shot answer a model could decompose it into multiple sub-questions, address each separately and aggregate the overall decision. Each of these steps (how to decompose? what to decompose to? etc.) could themselves be subjects of research.

C.3 Task-independent "Fairness". Many decision-making systems may carry biases and inequities. And with the increased reliance on computing technologies, we are running the danger of perpetuating their biases. Within NLP communities, there has been efforts to address bias (about gender, race, etc.), often with approaches that are limited to a certain representation or task. However, many are shown to be far-from-enough (11). Ideally we want to be able to bake-in high-level moral principles into the model — principles that are often pretty general and shared across tasks. While we, as a society, have relatively defined set of ethical principles, these considerations do not easily translate to the decisions made by algorithms. We need technologies that could better incorporate ethical considerations, ideally, with designs that involve transparency and explainability (§C.2) as validation of fairness.

C.4 Understanding "Sources" and their Ideologies. How can we help humans (§B) without understanding what is happening in their mind? Humans often act based on a set of beliefs that are shaped by their background. Beyond individuals, understanding of ideologies could be applied to any source of information (e.g., media outlets) to better understand the ideas and intents governing them. Writers, newspapers, TV channels are all sources of information that span continuum of ideologies about different topics and issues. We can use language technologies to better understand such beliefs and use this ability to better communicate with them.

D Funding the Research Plan

Bringing in external funding is an important step to maintaining a productive research group. In the past I have gained experience in writing research proposals by contributing to multiple small-scale grants. For instance, I spearheaded drafting grants for computational resources — a 15*k* AWS gift through Wharton Venture Initiation Program (2017-2020) and 20*k* Google Cloud Educational Research Grant (2019).

In terms of future funding programs, I am fortunate to be working in AI/NLP at a time when there is much interest. Major companies (Google, Microsoft, etc) have created joint initiatives to fund academic research. Earlier this year, the White House initiated an AI-focused program to directs Federal agencies to pursue well-defined targets. An example is DARPA's "AI Next", $2 billion program, which focuses on aspects like *explainability* (§C.2) and *common sense reasoning*. DARPA also has programs for other aspects of AI, like *fairness & AI* (§C.3) and *reducing the reliance on annotated data*. Other agencies like NSF, NIH, etc, similar programs as well.

References

[1] S. Chen, **D. Khashabi**, C. Callison-Burch, and D. Roth. Perspectroscope: A window to the world of diverse perspectives. In *ACL - demos*, 2019.

[2] S. Chen, **D. Khashabi**, W. Yin, C. Callison-Burch, and D. Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. In *NAACL*, 2019.

[3] P. Clark and O. Etzioni. My computer is an honor student - but how intelligent is it? standardized tests as a measure of AI. *AI Magazine*, 2016.

[4] John Corner. Fake news, post-truth and media–political change, 2017.

[5] N. DiFonzo. The echo-chamber effect. *New York Times*, 22, 2011.

[6] **D. Khashabi**, T. Khot, A. Sabharwal, P. Clark, O. Etzioni, and D. Roth. Question answering via integer programming over semi-structured knowledge. In *IJCAI*, 2016.

[7] **D. Khashabi**, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *NAACL*, 2018.

[8] **D. Khashabi**, T. Khot, A. Sabharwal, and D. Roth. Question answering as global reasoning over semantic abstractions. In *AAAI*, 2018.

[9] **D. Khashabi**, E. Sadeqi-Azer, T. Khot, A. Sabharwal, and D. Roth. On the capabilities and limitations of reasoning for natural language understanding. *arXiv:1901.02522*, 2019.

[10] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *EMNLP*, 2019.

[11] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL*, 2019.

[12] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N Smith. Annotation artifacts in natural language inference data. *NAACL*, 2018.

[13] T. Khot, A. Sabharwal, and P. Clark. Answering complex questions using open information extraction. *ACL*, 2017.

[14] L. Orman. Fighting information pollution with decision support systems. *Journal of management information systems*, 1(2):64–71, 1984.

[15] H. Trivedi, H. Kwon, T. Khot, A. Sabharwal, and N. Balasubramanian. Repurposing entailment for multi-hop question answering tasks. In *NAACL*, pages 2948–2958, 2019.

[16] B. Zhou, **D. Khashabi**, Q. Ning, and D. Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*, 2019.