Introduction

The goal of my research is to develop natural language understanding algorithms, that is, algorithms that induce a representation of meaning from natural language, and allow machines to **understand** text and **reason** over it. I focus on developing methods that **learn** to predict such meaning representations from data, rather than hand-coding meaning translation rules, as is common for instance in compilers. My algorithms take advantage of the rich structure of language to produce more accurate models, while using efficient inference procedures that can scale to large web-scale datasets.

The role of natural language understanding systems has risen dramatically in the last few decades (e.g., Apple's Siri, machine translation, Facebook's graph search), due to the explosion in the quantity of textual as well as non-textual information. While the amounts of available textual data are staggering, access to this body of knowledge is severely limited by the complexities of natural language. The broad goal of my work is to address the challenges posed by natural language understanding, namely, (a) *variability*: the same meaning can be expressed in multiple ways, and (b) *ambiguity*: the same language utterance can have multiple meanings.

Developing systems that understand natural language has the potential to revolutionize the way we interact with these vast amounts of data. Imagine driving your car and asking your mobile device to "find the cheapest gas station on Route 2 near Netanya". I work on methods that attempt to translate such input questions into a formal query (like an SQL query) that can be executed against a knowledge base (KB) to produce an answer. In addition, I develop methods that automatically find the knowledge necessary to answer such questions by structuring information retrieved from text. My research in natural language understanding enables applications such as dialogue systems, information extraction, question answering, and more.

I have worked in three main areas of natural language understanding: Textual Entailment, Semantic Parsing, and Machine Reading. My research has been recognized by several awards including ACL 2011 best student authored paper, ACL 2013 best paper runner-up, ACL 2014 best paper honorable mention, and EMNLP 2014 best paper award.

Textual Entailment Graphs

One of the fundamental challenges in developing language understanding systems is the problem of *language variability*: each target meaning can be expressed in natural language in a myriad of ways. Suppose we would like to answer the question *"Where is Ebola common?"*. We can infer the answer from the sentence *"Ebola is epidemic in Liberia."*, but for that we must know that the meaning of *"common"* can be inferred from the meaning of *"epidemic"*. In my PhD, I worked on automatically learning a large knowledge base (KB) of textual inference rules such as '*X epidemic in Y* $\Rightarrow$ *X common in Y*'.

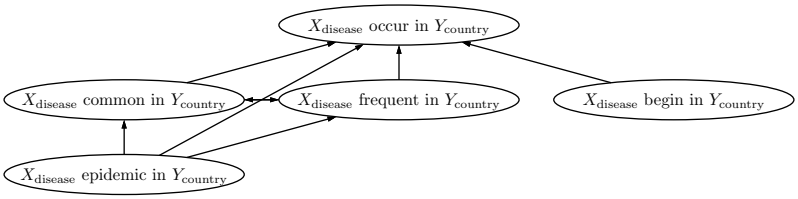

Figure 1: A fragment from a (transitive) entailment graph.

I cast the problem of learning inference rules as the problem of learning *entailment graphs* – graphs where nodes describe natural language predicates and edges represent inference or entailment relations (Figure 1). One of the fundamental properties of the entailment relation is that it is *transitive*, i.e., for any nodes x, y, z in the graph, $x \Rightarrow y$ and $y \Rightarrow z$ implies $x \Rightarrow z$. The goal is then to find the graph that maximizes the sum of

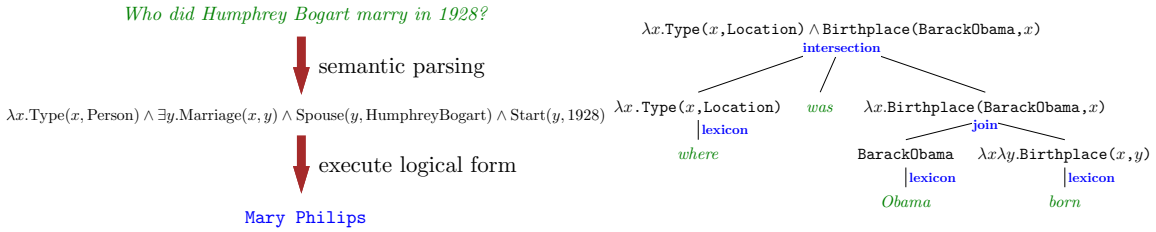

Figure 2: Left: Parsing language into an executable lambda calculus logical form. Right: A parse tree for the utterance *"Where was Obama born"*.

edge scores, under the structural constraint of transitivity.

I showed that finding the best entailment graph is NP-hard, and presented an (exponential) algorithm that, given a complete graph with weights on edges, selects the best subset of edges respecting transitivity. This algorithm formulated the problem as an Integer Linear Program (ILP) and solved it with an ILP solver (ACL 2010). To scale to large graphs, I subsequently developed a heuristic algorithm for maximizing the ILP objective, which is quadratic in the number of graph nodes (ACL 2011, ACL 2012). The algorithm is based on the observation that entailment graphs tend to be nearly "forest-reducible", i.e., they can be deterministically transformed into a forest. By restricting the hypothesis space to forest-reducible graphs I was able to design efficient inference algorithms that scale to graphs with tens of thousands of nodes. These graphs resulted in higher rule accuracy compared to previous state-of-the-art methods.

Semantic Parsing on Large Knowledge Bases

The goal of semantic parsing is to map natural language utterances to logical forms that can be executed against a knowledge base (Figure 2, left). This is a key technology for conversational interfaces on mobile devices, where users can control devices and obtain answers for questions using natural language. During my postdoctoral fellowship, I worked on developing semantic parsers that can scale to large knowledge bases, containing billions of facts. I have made the following contributions to this field:

Modeling language variability: As in textual entailment, one of the major challenges in scaling semantic parsers to large KBs is language variability, in this case handling the many ways in which phrases (*"born"*, *"native"*, etc.) map to logical predicates ('PlaceOfBirth'). I developed (EMNLP 2013) an unsupervised alignment algorithm that aligns millions of web sentences such as *"President Obama comes from Honolulu"* to millions of KB facts such as 'PlaceOfBirth(BarackObama,Honolulu)'. The algorithm outputs a lexicon that provides a mapping from phrases to predicates. I subsequently proposed (ACL 2014) an algorithm that learns to paraphrase questions in a way that simplifies finding the correct logical form. For example, paraphrasing the question *"Where is Obama from?"* to *"What is the place of birth of Barack Obama?"*. This is desirable since data for mapping phrases to KB predicates is limited, whereas paraphrase systems can be learned from the vast amounts of available textual data, in methods similar to learning of textual entailment graphs.

Learning to parse efficiently: Traditional semantic parsers use dynamic programming algorithms (such as CKY) that incrementally build parse trees (Figure 2, right) for longer and longer phrases. This is inefficient because all parse trees for phrases of length n must be generated before moving on to phrases of length $n + 1$. I developed an adaptive semantic parser that generates probable parse trees of any length before improbable ones, and thus substantially reduced computation. Crucially, the parser must learn to identify what actions are probable at each point of the computation. For example, the parser must learn that mapping the phrase *"Obama"* to the entity representing Barack Obama is more likely than to the entity representing the city Obama in Japan. I developed a novel algorithm for training semantic parsers

in a reinforcement learning framework and showed that the parser can learn to identify probable parsing actions, and consequently improve parsing speed without losing accuracy.

Our semantic parser obtains state-of-the-art results on multiple datasets, has been released in an open source package, and has been downloaded and utilized by groups from MIT, University of Washington, University of Edinburgh, and Xerox among others.

Machine Reading of Biological Processes

One of the higher-level goals of text understanding, often termed *machine reading*, is to develop systems that learn about the world by reading textbooks, just as children do in school. This is a much harder task than extracting simple facts from a large and highly redundant corpus such as the web. Towards accomplishing this goal, I developed a system that learns to read descriptions of biological processes from a biology textbook, and predict a meaning representation for them. The system demonstrates its understanding by answering complex questions about the described process. I used the biology domain in this work, but our approach can be extended to other domains in which processes are prevalent, like chemistry, economics, or even process descriptions on the web, such as on wikiHow.

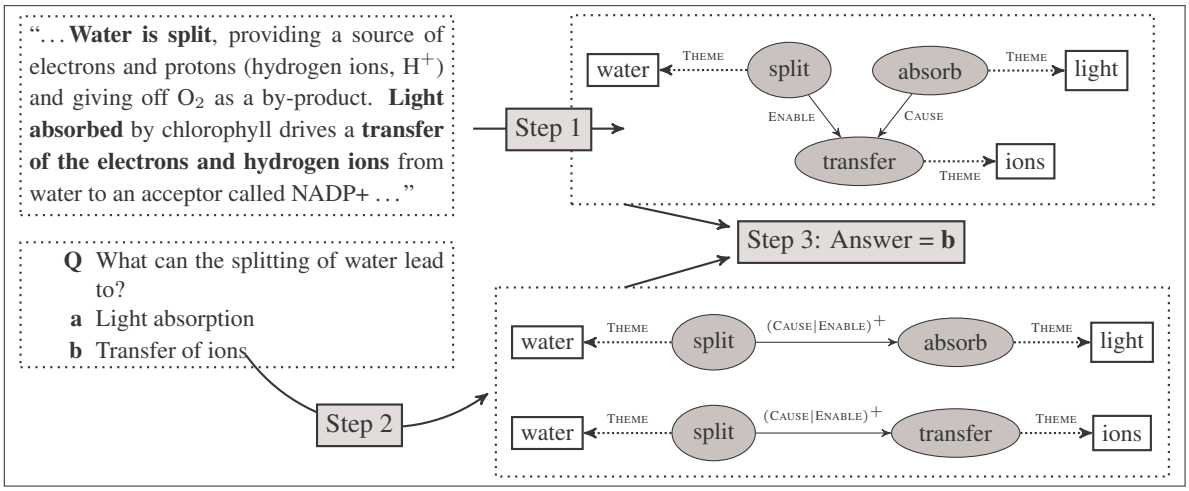

Figure 3: Machine reading of biolgical processes. First, we predict a structure from the input paragraph. Second, we map the question paired with each answer into a query that will be answered using the structure. Last, the two queries are executed against the structure, and a final answer is returned.

I modeled a process as a graph that describes relations between events and entities, and trained a structured predictor to predict process graphs from textual descriptions (Figure 3). Just like entailment graphs exhibit structural regularities, prior knowledge about the structure of processes can also improve the accuracy of our predictor. I used structural constraints, such as *graph connectivity*, which can be formulated as a linear inequality in an ILP (EMNLP 2013). Then, given a multiple-choice question about the process, the system treats the predicted structure as a small KB, and parses each possible answer to a query that is executed against the KB to determine the correct answer (EMNLP 2014). I have shown that generating a deep meaning representation for the processes improves question answering performance compared to shallower methods.

Future Research Directions

This is an exciting time to work on natural language understanding, a field that lies at the intersection of language, machine learning and logic. My long term research goal is to develop general language understanding algorithms that will enable easy language-based access to the plethora of textual information surrounding us. Some promising future research directions are:

Interactive learning: To evaluate my research, I have been involved in creating various datasets through crowdsourcing. While crowdsourcing eases the burden of annotating data, generating structured annotations is still an expensive and time-consuming endeavor. I would like to work on learning algorithms that do not require explicit data annotation and can instead bootstrap through live interaction with users.

Discourse and context: To provide a general model for language, we must take into account textual and non-textual context, including discourse and pragmatics. For instance, in my work on modeling biological processes, we handled paragraphs of text, but ideally we would like to incrementally read the entire textbook in a way that interpretation of later chapters depends on learning of earlier ones. I plan to work on modeling and learning from context and discourse, where information can propagate between higher-level units of natural language.

Non-executable semantic parsing: In my current work on semantic parsing, supervision is obtained by executing logical forms and comparing the result against a reference denotation (answer). However, in this setup, language must be limited to utterances that are represented in the target knowledge base (KB). An important future direction is to allow parsing of utterances that are not covered by the KB. To this end, one could automatically populate KBs on the fly from unstructured web text. Another direction is to avoid grounding in a KB, and instead obtain supervision by using the output meaning representation in an application that can pass feedback back to the learner. This could substantially increase the scope of semantic parsing to domains where no KB is available.

Symbolic and distributional representations of language: Recently, learned low-dimensional embeddings of language (vectors) had considerable success in NLP, as they allow the learning of representations that exhibit good generalization. In fact, our paraphrase model for semantic parsing employed such a representation. This development highlights the long-standing goal of combining symbolic representations such as first-order logic with distributional representations. I plan to examine the ways in which distributional methods can be combined with logical forms to produce learnable high-coverage meaning representations.

NLP in the service of science: I believe that research in NLP can provide tools and methods that will be useful in other scientific fields such as education, history, and biology (where I have worked as previously mentioned). I would like to work on NLP applications that advance research in other fields, and collaborate with researchers who require advanced semantic text processing.

Technological advances in recent years, and the increasing information load in modern life has made natural language processing applications imperative for structuring and accessing the large quantities of available textual knowledge. I am thrilled to be part of a field that has potential for such a high impact on society, and look forward to working on problems that will improve our understanding of language, and extend the applicability of NLP research.