Compositional and Modular Models for Reasoning over Text

Ph.D. Thesis Proposal

Nitish Gupta
Department of Computer and Information Science
University of Pennsylvania
nitishg@seas.upenn.edu

March 3, 2020

Ph.D. Thesis Advisor:
Prof. Dan Roth

Ph.D. Thesis Committee:
Prof. Mitch Marcus (Chair)
Prof. Lyle Ungar
Prof. Chris Callison-Burch
Prof. Luke Zettlemoyer

Abstract

In the last decade, deep artificial neural network models have become ubiquitous and have shown to achieve surprisingly exceptional performance on various natural language processing tasks. Despite such successes, several studies have shown that these models fail embarrassingly on seemingly trivial problems. The black-box nature of such models makes it difficult to interpret and debug their decision making process. In this thesis, I focus on developing modular and compositional models that reason over natural language, specifically models that read and answer questions against text as context. My work focuses on models that provide an understanding of the question semantics in terms of a formal executable parse which is composed of learnable modules that can perform reasoning over open domain text. Such models are desirable for various reasons – (a) being inherently compositional in nature, such models should be better able to capture the compositional nature of language and reasoning in general, which should result in accurate models, (b) the structured parse of the question and the outputs of intermediate modules make the model's decision making process interpretable and debuggable, (c) the modular nature of the model allows for transfer of supervision and reasoning capability across various domains and tasks. Until now, we have shown how models that perform natural language and symbolic reasoning over text can be designed and trained using end-goal supervision [2, 3]. We also showed that the use of auxiliary losses and external supervision leads to better performance. While such compositional models should be interpretable, we have also shown that it is difficult to achieve interpretability when trained in an end-to-end manner [1]. We have proposed a systematic and quantitative evaluation of interpretability and introduced ways to

achieve it. Going forward, I am focussing on two main directions – (a) understanding and improving systematic generalization in such models, and (b) extending such models to achieve transfer of reasoning capabilities across domains and tasks by sharing modules and supervision. The biggest challenge in pursuing this direction, and also what excites me the most, is trying to find a good middle-ground between the neatness of formal logic and fuzziness of reasoning over natural language.

Relevant Publications

The following publications contain work related to this thesis proposal:

1. (*) Sanjay Subramanian*, Ben Bogin*, Nitish Gupta*, Tomer Wolfson, Sameer Singh, Jonathan Berant and Matt Gardner. Achieving Interpretability in Compositional Neural Networks. *In submission (ACL 2020)*.

2. Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh and Matt Gardner. Neural Module Networks for Reasoning over Text. In *ICLR*, April 2020.

3. Nitish Gupta and Mike Lewis. Neural Compositional Denotational Semantics for Question Answering. In *EMNLP*, November 2018.

4. Shyam Upadhyay, Nitish Gupta and Dan Roth. Joint Multilingual Supervision for Cross-lingual Entity Linking. In *EMNLP*, November 2018.

5. Nitish Gupta, Sameer Singh and Dan Roth. Entity Linking via Joint Encoding of Types, Descriptions, and Context. In *EMNLP*, September 2017.

Contents

1 Introduction

We, as humans who understand how the world functions and ubiquitously use language to communicate, are able to seamlessly reason about natural language text and answer questions about the world. As a research community working towards achieving artificial intelligence, we would like to develop agents that are able to understand natural language and perform reasoning in a manner humans do. For example, we would like machines to be able to answer questions such as *"Which country has the highest per capita carbon dioxide emission?"* and *"What was the longest gap between two Radiohead albums?"*. Answering such questions goes beyond shallow lexical understanding of words; it requires an agent to understand concepts and events described in text, associate properties with them, and perform quantitative reasoning, as described linguistically by *highest* and *longest gap*. In this thesis, we aim to develop models that are able to perform *question answering*, i.e., map a natural language question to an answer given relevant context in a textual form.

Consider the question *"Which country has the highest per capita carbon dioxide emission?"* — an agent would ideally decompose this problem into multiple interrelated but simpler problems and answer the orginal question by solving these sub-problems. The most likely plan an agent could follow is to locate the *"countries"* mentioned in the context, for each one of them find their respective "per capita carbon dioxide emission", find the highest value amongst these and provide as answer the country with this emission value. Developing models to answer such questions, even against a single paragraph of text as context, poses a variety of challenges – primarily a system needs to be able to understand the compositional nature of language and reasoning in general, specifically the question in this case. Furthermore, a system needs to understand the concept of *"countries"* and locate its instantiations in text, it needs to tackle various linguistic variations in which "per capita carbon dioxide emission" might be mentioned, and the system should also be able to perform symbolic reasoning required to perform the *highest* operation. Similarly, the question *"What was the longest gap between two Radiohead albums?"* is underspecified and poses different challenges; amongst many, it requires the system to infer that the linguistic construction *"longest gap"* in the context of two *albums* refers to a time-span measured in years where these years are the *release dates* of *Radiohead albums*.

Previous research to solve such problems can broadly be classified into three threads, semantic parsing, machine reading comprehension, and neural module networks. Semantic parsing, rooted in formal semantics, aims to map a natural language utterance (e.g. question, instruction, etc.) to a logical meaning representation. In the context of question answering, this meaning representation is usually an *executable logical form*, that can be thought of as a program, which can be executed against some representation of the world to get the desired output. Explicit modeling of *compositionality* in the meaning representation makes semantic parsing a desirable approach to solve such problems. In the context of machine learning, the logical form provides a rationale to the final outcome predicted by the model which again makes semantic parsing a good choice to develop such question answering systems. On the other hand, semantic parsing bypasses the imporant questions of learning how to represent the world (context), and hence its usage is limited to modalities where execution can be determinisitically defined. For example, to answer questions against structured databases. **TODO: < −− SP assumes an unambiguous representation of the world where**

logical forms can be deterministically executed. In our case, where the context (world) where programs are executed is itself NL, it is challenging and requires learning a representation. Why this represenation cannot be a logical MR is because it is lossy and we assume that the questions we will tackle are much more restrictive than the contexts we will be encountering. Over the last decade, with the advent of large-scale neural network models for natural language processing, black-box neural models for question answering have emerged **(missing citations: rc models)**. Such models exploit the expressive representational capacity of neural networks to learn a "*meaning representation*" of language expressed as continuous vectors and provides an answer to the question without resorting to explicit compositional semantics. Such models have shown extremely good performance on standard benchmarks for machine reading comprehension, but there have also been several studies showing brittleness of these approaches **(missing citations: adversarial squad, pathologies, sears, etc.)**. Furthermore, the "black-box" nature of such approaches makes it difficult to interpret the model's decision making process at any scale. One key issue with such approaches is that they treat question answering as a problem of learning a single function to map questions and contexts to answers. However, as decribed in the earlier, it is perhaps useful to treat question answering as a multitask problem where each instance requires solving several interrelated problems. Neural module networks **(missing citations: NMNs; Andreas 2016)** carries this intuition forward and marries the approaches of formal semantic parsing with representation learning. It is a class of machine learning model where an utterance is mapped to a logical form, but where this logical form is composed of predicates that are not predetermined functions but rather functions with learnable parameters (or modules). The idea is that this set of primitive functions are learned to solve basic tasks of understanding the context which can be composed to perform higher-order reasoning. This approach decomposes the problem of learning a highly complex mapping function into the problem of learning an explicit meaning representation of the utterance and learning multiple primitive task predictors. While extremely promising, in practice this approach has mainly been applied to visual question answering in synthetic domains.

In this thesis, we aim to borrow ideas from these research directions and investigate how to further develop models for reasoning over text as context. Specifically, we would like the models we design to have the following desiderata; (a) the model should explicitly model compositionality in language and reasoning, i.e., the model structure should imitate the linguistic and reasoning structure closely, (b) the model's decision making process should be interpretable to some some level; this allows for understanding the model behavior and opens up possibilities for debugging, (c) the model should be modular, i.e. composed of operators that are resuable; this should allow for transfer of supervision and primitive reasoning capability across various domains and tasks. In Chaper 1, we present a neural module network (NMN) for answering compositional questions that require multiple steps of reasoning against text as context. We introduce modules that are capable of performing both shallow reasoning over a paragraph of text and symbolic reasoning (such as arithmetic, sorting, counting) over numbers and dates in a probabilistic and differentiable manner. The model we present fulfils all desiderata as explained above; but learning such a model using only weak question-answer supervision is extremely challenging. We additionally show how problem decomposition allows for using auxiliary objectives to aid learning. While such models are

inherently interpretable by the means of the question program and the intermediate outputs of the predicates in the logical form, we show in Chapter 2 that learning from end-goal supervision does not guarantee that the modules outputs are faithful to the logical meaning representation of the question. We find that due to the extreme expressivity of neural models, the modules do not learn their intended behavior when the only learning supervision is from the end-task. We outline few methods to alleviate this issue; providing auxiliary supervision for module outputs, designing the formal language such that expected linguistic phenonmena have corresponding semantic predicates, and providing the correct inductive bias to the model through module architecture design. Ultimately,

Some arbitrary cite (Grycner et al., 2015) and another arbitrary newcite Grycner et al., 2015. King of arbit.

A Appendix

A.1 Subsection R

A.2 Subsection H

References

Grycner, A., Weikum, G., Pujara, J., Foulds, J. R., and Getoor, L. (2015). Relly: Inferring hypernym relationships between relational phrases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 971–981. Association for Computational Linguistics.