Systematic language understanding: a study on the capabilities and limits of language understanding by modern neural networks

Koustuv Sinha
Ph.D. Proposal Document

September 2021

1 Introduction

Language allows us to express and comprehend a vast variety of novel thoughts and ideas. Through language, humans exhibit higher-order reasoning and comprehension. Thus, to develop models which mimic human-like reasoning, a principled focus in computer science research is to develop models which understand and reason on natural language. To foster research in developing such state-of-the-art natural language understanding (NLU) models, several datasets and tasks on reading comprehension have been proposed in recent literature. These include tasks such as question answering (QA), natural language inference (NLI), commonsense reasoning to name a few. Over the last decade, several advancements have been made to develop such models, the most successful ones till date involve deep neural models, especially Transformers , a class of multi-head self-attention models. Since its introduction in 2017, Transformer-based models have achieved impressive results on numerous benchmarks and datasets, with BERT being one of the most popular instantiation of the same. Using a technique known as "pre-training", Transformer-based models are first trained to replicate massive corpus of text. Through this kind of unsupervised training, the models learn and tune their millions and billions of parameters, and using which they solve NLU datasets with surprising, near-human efficiency .

While Transformer-based models excel in these datasets, it is less clear why do they work so well. Due to the sheer amount of overparameterization, direct inspection of the inner workings of these models are limited. Thus, various research have been conducted by using auxilliary tasks and probing functions to understand the reasoning processes employed by these models [3]. It has been claimed in the literature that BERT embeddings contain syntactic information about a given sentence, to the extent that the model may internally perform several natural language processing pipeline steps, involving parts-of-speech tagging, entity recognition etc . BERT has also been credited to acquire some level of semantic understanding , and contains relevant information about relations and world knowledge . All of these results indicate to the fact that purely pre-training with massive overparameterized models and large corpora might just be the perfect roadmap to achieve "human-like" reasoning capabilities.

On the other hand, there have been growing concerns regarding the ability of these NLU models to understand language in a "systematic" and robust way. The phenomenon of *systematicity*, widely studied in the cognitive sciences, refers to the fact that lexical units such as words make consistent contributions to the meaning of the sentences in which they appear [Fodor]. As an illustration, they provide an example that all English speakers who understand the sentence "John loves the girl" should also understand the phrase "the girl loves John". In case of NLU tasks, this accounts to model being consistent in understanding novel compositions of existing, learned words or phrases. However,

there is growing evidence in literature which highlight the brittleness of NLU systems to such adversarial examples . More so, there is strong evidence that state-of-the-art NLU models tend to exploit statistical artifacts in datasets, rather than exhibiting true reasoning and generalization capabilities .

In view of the positive and negative evidences towards Transformers acquiring "human-like" natural language understanding capacity, it is very important that we take a step back and carefully examine the reasoning processes of the NLU models in the view of systematicity and robustness. Since these NLU models are now being deployed in production and decision making systems, it is even more prudent to test the models towards systematic understanding in order to avoid catastrophic scenarios. In this proposal, I thus discuss my work till now in my doctoral studies to understand the limits of systematic and robust natural language understanding of NLU models. Concretely, first I discuss our proposed systematicity tests on artificial and natural languages by using first-order logic (FOL), and what we learned from the model using such tests. This involves the following two papers:

- *CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text*, published at EMNLP 2019 (Oral presentation)
- *Probing Linguistic Systematicity* [1], published at ACL 2020

Secondly, I discuss our work on understanding the limits of systematicity of NLU models by subjecting these models to scrambled word order sentences, involving the following two papers:

- *UnNatural Language Inference*, published at ACL 2021 (Oral presentation, Outstanding paper award) [5]
- *Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little*, submitted to EMNLP 2021 [4]

This document does not discuss concurrent works on understanding systematic reasoning for proof generation [1] (published at NeurIPS 2020), proposed dataset on systematic reasoning on graph neural networks [8], or an unreferenced automatic dialog evaluation framework [6] (published at ACL 2020) conducted during my doctoral studies.

2 Background

2.1 Language Models

2.2 The rise of pre-training

3 Contribution 1: Investigating systematicity of NLU models using first order logic

In this section, we first talk about our paper, CLUTRR [7]. Then, we talk about Probing Linguistic Systematicity [2] paper.

4 Contribution 2: Probing systematicity of pre-trained models using word order

Here, we talk about two related contributions, UnNatural Language Inference [5] and pretraining with unnatural languages [4].

[1] Work done as second author.

5 Future Work & Timeline

5.1 Unsupervised syntax learning by mutually exclusive training using word order

5.2 Entity agnostic training for better systematic generalization

5.3 Timeline

Thesis preparation and submission Expected defense date Fall 2022

References

[1] N. Gontier, K. Sinha, S. Reddy, and C. Pal. Measuring Systematic Generalization in Neural Proof Generation with Transformers. *arXiv:2009.14786 [cs, stat]*, Oct. 2020.

[2] E. Goodwin, K. Sinha, and T. J. O'Donnell. Probing Linguistic Systematicity. *arXiv:2005.04315 [cs]*, Aug. 2020.

[3] A. Rogers, O. Kovaleva, and A. Rumshisky. A Primer in BERTology: What we know about how BERT works. *arXiv:2002.12327 [cs]*, Nov. 2020.

[4] K. Sinha, R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. *arXiv:2104.06644 [cs]*, Apr. 2021.

[5] K. Sinha, P. Parthasarathi, J. Pineau, and A. Williams. UnNatural Language Inference. *arXiv:2101.00010 [cs]*, June 2021.

[6] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau. Learning an Unreferenced Metric for Online Dialogue Evaluation. *arXiv:2005.00583 [cs]*, May 2020.

[7] K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. *arXiv:1908.06177 [cs, stat]*, Sept. 2019.

[8] K. Sinha, S. Sodhani, J. Pineau, and W. L. Hamilton. Evaluating Logical Generalization in Graph Neural Networks. *arXiv:2003.06560 [cs, stat]*, Mar. 2020.