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 [10]. 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 Empirical Methods of Natural Language Processing (EMNLP) 2019 (Oral presentation) [15]
- *Probing Linguistic Systematicity* [1], published at Association for Computational Linguistics (ACL) 2020 [3]

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 Association for Computational Linguistics (ACL) 2021 (Oral presentation, Outstanding paper award) [13]
- *Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little*, published at Empirical Methods of Natural Language Processing (EMNLP) 2021 [12]

This document does not discuss concurrent works on understanding systematic reasoning for proof generation [2] (published at NeurIPS 2020), analyzing faithfulness and robustness in translations [9] (published at EMNLP 2021), proposed dataset on systematic reasoning on graph neural networks [16], or an unreferenced automatic dialog evaluation framework [14] (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 [15]. Then, we talk about Probing Linguistic Systematicity [3] paper.

[1] Work done as second author.

3.1 CLUTRR

3.1.1 Motivation

An important challenge in NLU is to develop benchmarks which can precisely test a model's capability for robust and systematic generalization. Ideally, we want language understanding systems that can not only answer questions and draw inferences from text, but that can also do so in a systematic, logical and robust way. While such reasoning capabilities are certainly required for many existing NLU tasks, most datasets combine several challenges of language understanding into one, such as co-reference/entity resolution, incorporating world knowledge, and semantic parsing - making it difficult to isolate and diagnose a model's capabilities for systematic generalization and robustness.

Thus, inspired by the classic AI challenge of inductive logic programming, we propose a semi-synthetic benchmark designed to explicitly test an NLU model's ability for systematic and robust logical generalization. Our benchmark suite - termed CLUTRR (Compositional Language Understanding with Text-based Relational Reasoning) - contains a large set of semi-synthetic stories involving hypothetical families. Given a story, the objective is to infer the relationship between two given characters in the story, whose relationship is not explicitly mentioned. To solve this task, a alearning agent must extract the relationships mentioned in the text, induce the logical rules governing the kinship relationships (e.g, the transitivity of the sibling relation), and use rules to infer the relationship between a given pair of entities.

3.1.2 Dataset Design

In order to design the CLUTRR benchmark, we build upon classic ILP task of inferring kinship relations. For example, given the facts that *"Alice is Bob's mother"* and *"Jim is Alice's father"*, one can infer with reasonable certainty that *"Jim is Bob's grandfather"*. While this example may appear trivial, it is challenging task to design models that can learn from data to *induce* the logical rules necessary to make such inferences, and it is even more challenging to design models that can systematically generalize by composing these induced rules. Thus, the core idea behind CLUTRR benchmark suite is the following: given a natural language story describing a set of kinship relations, the goal is to infer the relationship between two entities, whose relationship is *not* explicitly stated in the story. To generate these stories, we first deign a knowledge base (KB) with rules specifying how kinship relations resolve, and we use the following steps to create semi-synthetic stories based on this knowledge base:

- **Step 1.** Generate a random kinship graph that satisfies the rules in our KB.
- **Step 2.** Sample a target fact (i.e relation) to predict from the kinship graph
- **Step 3.** Apply backward chaining to sample a set of facts that can prove the target relation (and optionally sample a set of "distracting" or "irrelevant" noise facts)
- **Step 4.** Convert the sampled facts into a natural language story through pre-specified text templates and crowd-sourced paraphrasing.

3.1.3 Experiments and Discussion

3.2 Probing Linguistic Systematicity

3.2.1 Motivation

3.2.2 Experiments

3.2.3 Discussion

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

Here, we talk about two related contributions, UnNatural Language Inference [13] and pretraining with unnatural languages [12].

4.1 UnNatural Language Inference

4.2 Pre-training with Unnatural Languages

5 Future Work & Timeline

Until now, most of the work in my doctoral studies have been focused on developing methods to detect the issue of systematicity plaguing neural NLU models. To complete my thesis, I thereby plan to work towards a couple of methods to improve robustness and systematicity of NLU models.

5.1 Unsupervised syntax learning by mutually exclusive training using word order

In our prior work on word order (Section 4), we observed that NLU models are largely distributional - they understand the collection of words in a sentence but have limited understanding on the order of words. This poses a problem - representations of random permutations of the sentence having no grounded meaning will still be identified by the NLU models to contain syntactic and semantic information. Due to this distributional effect, we posit that syntax understanding of NLU models are still primitive, mostly restricted to higher order information. Thus, it is imperative to develop mechanisms to imbibe the required syntactical information within the sentence representation, such that it is systematic. One can use syntactic features such as dependency parses to imbibe information about syntax in the sentence representation using auxilliary supervison loss. In literature, such syntactical information has shown to be effective in downstream tasks, such as Relation Extraction (RE) [1], named entity recognition (NER) [5] and semantic role labeling (SRL) [18]. More recently, syntax trees are used during pre-training of Transformer based models to imbibe better syntactical information by early and late fusion [11]. However, such direct supervision models to imbibe syntactical information is difficult as it requires access to preferably human-annotated syntax parses of sentences, which raises questions on the viability of such approaches for real world applications. Even so, limited studies have been performed to investigate systematicity issues of those models trained with supplementary syntactical signal.

Therefore, we propose an alternate, unsupervised mechanism to imbibe syntax information within the sentence representations by leveraging word order. Concretely, we use an auxilliary objective to the model to recognize correct and incorrect permutations of a given sentence alongside the task objective. Now, in the strictest sense there is only one correct ordering of a sentence which conveys the intended meaning. However, natural language (English) allows for a degree of flexibility in word order. Thus, we plan to leverage the idea of *separable permutations* [17], where a subset of permutations can be treated as positive signal which can be reconstructed from the CCG parse of the given sentence. This auxilliary training loss could potentially inform the model to be systematic in understanding syntax, and thereby reduce the distributional, bag-of-words behavior of the encoder representations.

5.2 Nonsensical data augmentation for better systematic generalization

Systematic generalization is an issue which plagues many NLU tasks, in particular Natural Language Inference (NLI). Generalization to out-of-domain examples is poor [8], and it has been shown that these models leverage the statistical artifacts in NLI datasets, such as SNLI and MNLI [4]. One reason why models tend to overfit on the training data is the exposure bias to specific nouns/verbs/entities during training. When subjected to systematic stress test, the NLU models tend to be brittle as they fail to learn syntax of the training signal by fixating on the rare words and artifacts. Thus, we propose a dynamic data augmentation training scheme for NLU models where we repeat the training examples with word replacements from the same syntactic family. Overall, a sentence might lose its intended meaning (hence, “nonsensical”) - however if the same operation is conducted on the premise, the entailment logic remains unchanged. Concretely, given a lexical item in a sentence, we randomly replace that item with another belonging to the same syntactic family, equally in both premise and hypothesis sentences. For empirical reasons we will restrict this replacement to specific family of lexicons (proper nouns, verbs) which typically form as rare elements in the dataset. We also plan to include a probabilistic model for this replacement which replaces lexicons based on their corpus probability to ensure uniformity in training. By systematically replacing the lexicons in a different context one can potentially increase the training data to reduce exposure bias problem. Similar methods have been devised for mitigating gender bias previously in the literature with varying success [6].

5.3 Timeline

Unsupervised syntax learning by mutually exclusive training using word order (1) Investigate CCG parsing to generate separable permutations on the fly, (2) Representational analysis on separable and non separable permutations, (3) Train auxillary loss with either direct supervision or partial gradient based methods such as Meta Learning, (4) Write paper for ACL 2022 or TACL 2022

Nonsensical data augmentation for better systematic generalization (1) Investigate lexicon replacements by using syntactic parsers in a given dataset, (2) Analyze how the distribution of rare elements change in the training corpus by using this kind of replacement, (3) Test on out-of-domain data and NLI stress test sets, such as HANS [7], (4) Write paper for EMNLP 2022.

Thesis preparation and submission Expected defense date: Fall 2022

References

[1] K. Fundel, R. Küffner, and R. Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[2] 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.

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

[4] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation Artifacts in Natural Language Inference Data. *arXiv:1803.02324 [cs]*, Apr. 2018.

[5] Z. Jie and W. Lu. Dependency-guided lstm-crf for named entity recognition. *arXiv preprint arXiv:1909.10148*, 2019.

[6] R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel. It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution. *arXiv:1909.00871 [cs]*, Feb. 2020.

[7] R. T. McCoy, E. Pavlick, and T. Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv:1902.01007 [cs]*, June 2019.

[8] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. *arXiv:1910.14599 [cs]*, May 2020.

[9] P. Parthasarathi, K. Sinha, J. Pineau, and A. Williams. Sometimes We Want Translationese. *arXiv:2104.07623 [cs]*, Apr. 2021.

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

[11] D. S. Sachan, Y. Zhang, P. Qi, and W. Hamilton. Do Syntax Trees Help Pre-trained Transformers Extract Information? *arXiv:2008.09084 [cs]*, Jan. 2021.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] M. Stanojević and M. Steedman. Formal Basis of a Language Universal. *Computational Linguistics*, 47(1):9–42, Apr. 2021.

[18] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum. Linguistically-Informed Self-Attention for Semantic Role Labeling. *arXiv:1804.08199 [cs]*, Nov. 2018.