Metrics and Language in NLP Research

Rod Ali Mazloomi
NLP Reading Course Final Paper
Faculty of Information
University of Toronto

"When a measure becomes a target, it ceases to be a good measure"

Goodhart's Law
(Strathern, 1997)

I. ABSTRACT

In the last decade of Natural Language Processing (NLP) research, language models have consistently achieved state-of-the-art performance on many metrics, leading some to claim that these models are *understanding* and *comprehending* human language. Goodhart's Law states that "When a measure becomes a target, it ceases to be a good measure." I apply this to the field of NLP where metrics and benchmarks are used to evaluate the performance of language models on language tasks. In this short essay, I argue that these metrics can be misleading and lack validity. I draw from a growing body of research that demonstrates that language models are fragile and that they do not pick up on the semantic structure of language. I relate this to the concepts of external and construct validity, and I comment on the harmful social consequences. Finally, I offer two suggestions for improvement that I have garnered from work in the field. My motivation is to critically assess our evaluation methods in NLP such that we are better able to understand their strengths and limitations.

II. INTRODUCTION

The field of Natural Language Processing (NLP) in the last decade has seen advances in the ability of computational models to solve difficult language tasks. Advancements in deep learning have led to sophisticated models emerging. Multilayer perceptrons (MLP), Long-short term memory (LSTM) – a form of recurrent neural network (RNN), and most recently Transformers have all been pushing the performance on several evaluation metrics and benchmarks.

Google Brain in 2017 developed the Transformer architecture which utilizes a self-attention mechanism consisting of stacks of encoder and decoder modules (Vaswani et al., 2017). Based on this architecture, pretrained models such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) emerged. These pre-trained models were trained on large online corpora containing over millions of tokens. BERT was trained on BooksCorpus (800 million tokens) and the Wikipedia corpus (2,500 million tokens)

(Devlin et al., 2019). These models are very successful in NLP research because of their strengths with textual data. BERT achieved state-of-the-art performance on eleven language tasks, including the competitive Stanford Question Answering Dataset (SQuAD) benchmark (Devlin & Chang, n.d.). BERT achieved an F1-score of 93.16 on SQuAD, surpassing human performance of 91.22 by a difference of 1.94.

There is no doubt that the performance of these models is impressive. Surpassing human performance has generated much enthusiasm among researchers, leading some to claim that language models are able to *understand*, *comprehend* and represent the *semantical structure* (meaning) of language; examples of some of these claims are provided in section VI (Bender & Koller, 2020). These are large overclaims. In this short essay, I argue that the metrics used in NLP research can be misleading and lack validity. Language models surpassing human performance on a single metric does not imply that they contain the same capacities as human beings.

For the sake of this paper, I will use the term *language models* to refer to neural networks that are trained on language data and perform language tasks. *Benchmarks* are datasets that are used to evaluate the performance of language models on language tasks, and *metrics* are the quantitative value that represents this performance. Examples of benchmarks include SQuAD, and GLUE (General Language Understanding Evaluation) (Glue Benchmark, n.d.). Some metrics include F1-score for classification and BLEU (bilingual evaluation understudy) for translation (Papineni et al., 2001).

This paper is divided into the following sections: (III) discusses the fragility of language models and their challenges with generalizability, (IV) addresses that language models do not learn what we think they are learning, (V) connects these challenges to the ideas of external and construct validity from experimental design, (VI) discusses the negative social consequences of neglecting these challenges, and (VII) I will offer suggestions for improvement and (VIII) will be a brief conclusion.

III. LANGUAGE MODELS ARE FRAGILE

Language models are achieving high accuracy on multiple language tasks, including translation, question answering and inference. Despite this, there is a growing body of work that shows that language models are fragile when put into new contexts. This has led many researchers to construct new benchmarks that test the same language task and maintain human performance. They find that when language models are tested on these constructed datasets that their performance plummets significantly, sometimes below random guessing.

Language model accuracies have consistently increased on question answering tasks. Jia and Lang sought to answer whether this increase was due to *understanding* or by some other means. They explore the performance of some sixteen language models that performed well on SQuAD with an average F1-score of 75% (Jia & Liang, 2017). They develop an adversarial

evaluation scheme that inserts distraction sentences into the SQuAD benchmark that deteriorates language model performance yet maintains human performance. The adversarial evaluation scheme drops the performance of the language models to 36%. In certain conditions, the F1-score of four of the models dropped as low as 7%. They conclude by recommending their adversarial evaluation scheme for a more accurate evaluation of language models on question answering tasks.

Belinkov and Bisk evaluate the performance of language models in machine translation (Belinkov & Bisk, 2018). Although language models had an impressive performance on translation tasks, they find that they are "very brittle and easily falter when presented with noisy data" (Belinkov & Bisk, 2018, p. 1). In their study, they look at three types of noisy data: (1) random permutation of words, (2) swapping pairs of adjacent letters, and (3) natural human errors. They find that when text contains noise performance drops significantly. Drawing from psychological research, they compare the performance of language models to human performance. They indicate that humans have "robust language processing systems" (Belinkov & Bisk, 2018, p. 1) that can overcome noise, such as typos, misspellings, and missing characters. Noise is very common in written language, and it presents a difficult obstacle in the path of language models in understanding language.

Zellers and colleagues were quite surprised to see that language models were achieving near human performance on language inference (Zellers et al., 2019). They constructed a new benchmark dataset to evaluate language model's performance on language inference. They used adversarial filtering, "a data collection paradigm wherein a series of discriminators iteratively select an adversarial set of machine-generated wrong answers" (Zellers et al., 2019, p. 1). They found that while the questions were simple for humans with a >95% accuracy, language models struggled with an accuracy of <48%. They conclude by suggesting that benchmarks should continually co-evolve with state-of-the-art models.

The generalization of language models to novel contexts remains a challenge and an active area of research. We see the performance of language models drop when evaluated in a context other than the one they were trained. Would we not expect performance to maintain if language models were representing the semantical structure of language? Adversarial methods – techniques to attempt to exploit models' performance to gain a deeper understanding of a model's structure – is a promising avenue to improve generalizability. As our models evolve, so must our benchmarks and metrics that are used to evaluate their performance on language tasks. This ensures that we understand the power of language models, and just as important, their limitations.

IV. LANGUAGE MODELS ARE NOT LEARNING WHAT WE THINK

The inherent complexity of language models presents a challenge to understand them. This has led to them being referred to as a *black box*, where we are able to observe their input and output, but not the process to which the input was mapped to the output. High performance

on metrics can therefore be misleading if the process is unclear. Achieving high performance on a language task is not sufficient; it is just as important to understand *how* that model achieved its performance. Several studies have explored the inner workings of language models.

Niven and Kao are surprised to see that BERT achieves 77% accuracy on the Argument Reasoning Comprehension Task (ARCT), just three points below the average human baseline (Niven & Kao, 2019). They ask, "This motivates the question: what has BERT learned about argument comprehension?" (Niven & Kao, 2019, p. 1). In ARCT, the language model is provided with *claims* and *reasons* and must discover the *warrant* that draws the inference between the claim and reason. They investigate BERT's decision-making using a probing experiment, designed to isolate the effects of cue words. They find that "BERT exploits the presence of cue words in the warrant, especially 'not'" (Niven & Kao, 2019, p. 2). Furthermore, they demonstrate that "BERT's surprising performance can be entirely accounted for in terms of exploiting spurious statistical cues" (Niven & Kao, 2019, p. 2). They do this by constructing an adversarial test set that negates the claim and label. They find that the performance of BERT drops to the equivalence of a random guess. Answering their initial question, "To answer our question in the introduction: BERT has learned nothing about argument comprehension" (Niven & Kao, 2019, p. 5). They suggest that their adversarial dataset should be adopted for future work to provide a more accurate evaluation of argument comprehension.

Musgrave and colleagues find three flaws in experimental methodology in numerous deep learning papers (Musgrave et al., 2020). They find unfair comparisons of model hyperparameters, accuracy metrics that are misleading, and training without a validation set. This leads them to conclude that "actual improvements over time have been marginal at best" (Musgrave et al., 2020, p. 1). They propose an evaluation protocol that addresses these problems.

Agrawal and colleagues explore the behaviour of language models on the task of Visual Question Answering (VQA) (Agrawal et al., 2016). Their analysis shows that models tend to "fail on sufficiently novel instances" (Agrawal et al., 2016, p. 1959), and they "jump to conclusions" (Agrawal et al., 2016, p. 1959) after listening to half the question. The deep learning models are "stubborn" (Agrawal et al., 2016, p. 1959), they pick on a few cues and do not change their answers across images.

Other lines of work show that semantics or the extension of words cannot be represented by language models and that these models are limited to form and statistical cues. Bender and Koller argue that language models have no *a priori* method of learning meaning or understanding natural language (Bender & Koller, 2020). They support their point by presenting several thought experiments that show that meaning cannot be derived by observing form alone. Likewise, Herbelot highlights the extension of the term as crucial to lexical meaning, "corpora fail to supply the information necessary to represent the extension of a term" (Herbelot, n.d., p. 1). Herbelot continues to say, "that an appropriate representation of lexical meaning requires information beyond what is provided by written text and that this can be problematic for lexical models which rely entirely on corpora" (Herbelot, n.d., p. 1).

Language models capture statistical cues and the syntax of language. Semantical structure and the extensional aspect of language remain out of the bounds of language models. This can be discouraging, but there is great usefulness in language models that represent syntax, such as spellchecking and translation systems. As Herbelot mentions, if we want to represent the lexical meaning of language, information is required *beyond* language.

V. EXTERNAL AND CONSTRUCT VALIDITY

In experimental design, the validity of an experiment is how you are measuring what you claim to be measuring (Validity, n.d.). There are multiple dimensions to validity. Two of which are relevant are *external validity* and *construct validity*. I argue that the challenges identified in the previous two sections are symptoms of weak external and construct validity.

External validity is whether the results of your experiment can be applied outside of the experimental setting. It assesses whether your findings can be applied to other contexts. There is not much purpose in a model that performs well in an experiment and not in any other setting. A goal of research is that our experiments will lead to knowledge that can be applied to the world – outside of the experimental setting. In the case of NLP research, external validity is synonymous with the generalizability of a model. It evaluates how well a model can be used in other contexts and maintain its performance. As we have seen external validity is a challenge in NLP research. Language models that achieve state-of-the-art performance on a benchmark in one context, can be reduced to random guessing when put in another (Niven & Kao, 2019). This tells us that our models are representing information from the experiment that is not transferable to other contexts. Humans are incredible at abstraction and functioning in environments that they are not well familiar with. If we are to claim that language models are *learning*, we must recognize the stark contrast between *human learning* and *machine learning*.

Another dimension of validity is construct validity. Construct validity focuses on what we *claim* to be measuring. A construct is a concept and construct validity is how well our measure is indicative of our concept. Construct validity can be straightforward, for example, measuring one's height with a measuring stick, or questionable, such as measuring an individual's intelligence using an IQ (intelligence quotient) test. Construct validity is of vital importance to ensure that we know what we are measuring. Construct validity of language models is how our metrics and benchmarks are indicative of the construct we are assessing. But what is the construct that our benchmarks are assessing in NLP? Some have claimed that benchmarks are assessing the ability to understand, comprehend and grasp semantics (Bender & Koller, 2020). As the literature in the previous section has shown, language models pick up on syntax and statistical cues (Niven & Kao, 2019). As researchers, if we think our language models are learning the semantics of language, our experiments suffer from weak construct validity.

VI. SOCIAL CONSEQUENCES OF IMPRECISE LANGUAGE

There exists a schism between the expectation and the reality of language models. This can be seen in the language researchers use in their work. Bender and Koller analyze how language is used in NLP research and find that terms such as *understanding,* and *comprehension* are frequently used (Bender & Koller, 2020). Three such sources from Bender and Koller, with emphasis added, include:

(1) "In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task." (Devlin et al., 2019, p. 4)
(2) Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension** . . . (Ohsugi et al., 2019, p. 1)
(3) "The surprisingly strong ability of these models to recall **factual knowledge** without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems." (Petroni et al., 2019, p. 1)

Such language is imprecise with our current understanding of how language models work. When the same vocabulary is used to describe human learning and machine learning misunderstandings are created. This leads to the propagation of false information and AI hype in the media (Bender & Koller, 2020). This imprecise language misleads the public and develops fictional ideas in the discourses of wider society.

Birhane and Dijk emphasize the dangers of techno-optimism and reductionism in the field of Artificial Intelligence (Birhane & Dijk, n.d.). They identify several ideas, such as General AI, the singularity and super-intelligence that are overemphasized in media. They say, "Romantic predictions like this, invariably envisioning breakthroughs some decades into the future, have been recurring since the earliest days of digital technology; all have proven empty" (Birhane & Dijk, n.d., para. 4). NLP has many challenges that exist today that should be the center of our discourse. Algorithmic discrimination, bias, and invasion of privacy to name a few. A realistic understanding and precise language are the first steps toward addressing weak construct validity.

The argument being made is not that *understanding* and *comprehension* are impossible, but rather that our language and understanding should describe language models as they exist *today* and not in some hypothetical future.

VII. SUGGESTIONS FOR IMPROVEMENT

Reviewing the research literature, I found roughly two overarching points that I will summarize here. The first is that language models and benchmarks should symbiotically coevolve. As our models become more accurate, so should our methods to evaluate them. This allows us to construct benchmarks that reflect a clear theoretical construct. Musgrave and associates note that "If proper machine learning practices are followed, and comparisons to prior work are

done in a fair manner, the results of future metric learning papers will better reflect reality and will be more likely to generalize to other high-impact areas like self-supervised learning." (Musgrave et al., 2020, p. 14). Adversarial datasets have been beneficial in much of the work that I have reviewed. These datasets allow us to dig deeper into how a model functions and test whether the model represents useful information about the world or spurious statistics that plague it with weak external validity. In their work, Niven and Kao conclude "As our learners get stronger, controlling for spurious statistics becomes more important to have confidence in their apparent performance. Taken with a growing body of previous work, our results indicate the need for further research into the extent of this problem in NLP more generally. The adversarial dataset should be adopted as the standard in future work on ARCT" (Niven & Kao, 2019, p. 5). When we have strong benchmarks, we are in the position to make advancements in the development of language model performance. This ensures that our models are learning what we want them to learn and that we evaluate them accordingly. Jia and Lang state "Progress on building systems that truly understand language is only possible if our evaluation metrics can distinguish real intelligent behaviour from shallow pattern matching" (Jia & Liang, 2017, p. 2029).

The second is that researchers should pay close attention to their language in their academic work. Using inaccurate language blurs the reality of how language models are functioning and performing. Bender and Koller state "If the highlighted terms are meant to describe human-analogous understanding, comprehension, or recall of factual knowledge, then these are gross overclaims. If, instead, they are intended as technical terms, they should be explicitly defined" (Bender & Koller, 2020, p. 2). Furthermore, the use of this language blurs our goals in the field of NLP. We see frequent advancements and the question arises, where are we headed? What is our goal? Is it to build language models that will translate a piece of text from one language to another, or to develop *Generalized Linguistic Intelligence* (Yogatama et al., 2019)? I will echo some of the qualities that are shared by Bender and Koller that are needed in our pursuit to refine our language. That we need to "be aware of the limitations of tasks" and "above all, cultivate humility towards language and ask top-down questions." (Bender & Koller, 2020, p. 8).

VIII. CONCLUSION

This essay draws inspiration from Goodhart's Law, that "When a measure becomes a target, it ceases to be a good measure." This work aims to bring to attention our metrics in NLP and how they can be misleading. I argued that this has to do with weak external and construct validity. I referenced a growing body of research that shows that language models are not generalizable and do not capture what we think they are. I commented on the harmful social impacts of neglecting weak validity. I concluded by offering two suggestions for improvement gathered from my review on this topic: that evaluation methods need to symbiotically evolve with our language models and that humility is needed in our language.

References

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the Behavior of Visual Question Answering Models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1955–1960. https://doi.org/10.18653/v1/D16-1203

Belinkov, Y., & Bisk, Y. (2018). Synthetic and Natural Noise Both Break Neural Machine Translation. *ArXiv:1711.02173 [Cs]*. http://arxiv.org/abs/1711.02173

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Birhane, A., & van Dijk, J. (n.d.). *Essay Philosophy & Culture*. 10.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

GLUE Benchmark. (n.d.). Retrieved April 21, 2022, from https://gluebenchmark.com/

Herbelot, A. (n.d.). *What is in a text, what isn't, and what this has to do with lexical semantics*. 6.

Jacob Devlin, & Ming-Wei Chang. (n.d.). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google AI Blog*. Retrieved April 20, 2022, from http://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Jia, R., & Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. https://doi.org/10.18653/v1/D17-1215

Musgrave, K., Belongie, S., & Lim, S.-N. (2020). A Metric Learning Reality Check. *ArXiv:2003.08505 [Cs]*. http://arxiv.org/abs/2003.08505

Niven, T., & Kao, H.-Y. (2019). Probing Neural Network Comprehension of Natural Language Arguments. *ArXiv:1907.07355 [Cs]*. http://arxiv.org/abs/1907.07355

Ohsugi, Y., Saito, I., Nishida, K., Asano, H., & Tomita, J. (2019). A Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension. *Proceedings of the First Workshop on NLP for Conversational AI*, 11–17. https://doi.org/10.18653/v1/W19-4102

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language Models as Knowledge Bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. https://doi.org/10.18653/v1/D19-1250

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

Validity (Internal, External, Construct)—StatsDirect. (n.d.). Retrieved April 19, 2022, from https://www.statsdirect.com/help/basics/validity.htm

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Yogatama, D., d'Autume, C. de M., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., & Blunsom, P. (2019). Learning and Evaluating General Linguistic Intelligence. *ArXiv:1901.11373 [Cs, Stat]*. http://arxiv.org/abs/1901.11373

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? *ArXiv:1905.07830 [Cs]*. http://arxiv.org/abs/1905.07830