Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/97/20241301

A Review of the Limitations of Language Models in NLP

Xuantong Zhang[1,3,*], Binyu Li[2,4]

[1]Capital Normal University High School, Beijing, 100048, China
[2]Tianjin Jiacheng Middle School, Tianjin, 300011, China

[3]xuantongzhang@ldy.edu.rs
[4]3197210463@qq.com
*corresponding author

Abstract. In the field of Natural Language Processing (NLP), large neural language models have successfully been applied to a variety of tasks, including machine translation and reading comprehension. These models often learn the structure of language (such as grammar and semantics) by predicting the next word or character. However, there's an overly optimistic assessment of the current state of natural language understanding, which assumes that models can handle text at a semantic level. This article primarily explores the successes of large neural language models (such as BERT and GPT-2) in numerous tasks that require the understanding of linguistic meaning in the domain of NLP. Language models trained purely on form cannot learn meaning, as understanding meaning involves the relationship between linguistic form and non-linguistic communicative intent. The article aims to guide scientific research in the field of Natural Language Understanding (NLU) by clearly distinguishing the concepts of form and meaning. The learning, generalization, and other capabilities of different models under various attributes show significant differences, suggesting that a combination of different models and properties can achieve unexpected results. The simple learning combinations of some models may offer new insights into the understanding of linguistic meaning by models.

Keywords: Natural language processing, LMs limitations, neural language model, formal language, linguistic meaning.

1. Introduction

In the expansive domain of Natural Language Processing (NLP), language models (LMs) serve as a pivotal force, persistently advancing the boundaries of language understanding and generation. However, as research delves deeper, it becomes increasingly evident that, despite remarkable strides in various tasks, these models face significant challenges in capturing the deeper meanings of language. This paper aims to delve into this reality by dissecting the limitations of language models in handling complex linguistic phenomena such as propositional logic, offering theoretical underpinnings and practical guidance for future model enhancements and developments. Language models, particularly the deep learning models that have emerged in recent years such as the Transformer and its variants, have markedly improved performance on NLP tasks. However, these models predominantly rely on statistical patterns gleaned from large corpora for learning and prediction, which restricts their ability to grasp the deeper semantics of language. Particularly in dealing with phenomena like propositional logic, which involves complex logical relationships and reasoning capabilities, existing models often fall short. Thus,

DOI: 10.54254/2755-2721/97/20241301

this study zeroes in on exploring whether language models, when only having access to linguistic form, can effectively capture and differentiate the logical symbols in language and the meaning they convey. The significance of this study lies in illuminating the limitations of current language models in understanding the deep semantics of language, impacting the advancement of NLP technology profoundly. Firstly, it prompts researchers to revisit the training methods and optimization objectives of existing models, pondering on how to design more rational training corpora and model architectures to enhance the models' ability to grasp the meaning of language. Secondly, this study provides new perspectives and directions for future research, such as how to improve models to better handle the combinatorial nature and logical structure of language, and how to incorporate external knowledge or common sense to augment the semantic understanding capabilities of models. Finally, through a thorough exploration of the limitations of language models, this study also opens up possibilities for interdisciplinary research, such as introducing theories and methods from linguistics, logic, cognitive science, and other fields into the realm of NLP, jointly advancing the innovative development of language understanding and generation technologies. This translation maintains the academic tone and structure of the original Chinese text, ensuring a comprehensive and precise translation for academic or technical contexts. It accurately conveys the complex ideas and implications of the research while adhering closely to the original content.

The relevant literature focuses on exploring the performance and limitations of Transformer and its associated models in NLP tasks. Literature [1] constructed a training corpus with semantic transparency, using Long short-term memory (LSTM) and Transformer models, and found their limitations in distinguishing logical symbols with different meanings. Literature [2] defines input-output pathways and state transition relationships, assessing the combinatorial performance of sequence-sequence models when dealing with complex environments. Literature [3] compares the generalization ability and learning mechanism of Transformer and LSTM, and finds that Transformer performs well on only a subset of regular languages, and its performance decreases with the increase of language complexity. Literature [4] proposes a new hard attention mechanism Unique Hard Attention (UHAT), and constructs the Boolean circuit to recognize AC language. It is found that UHAT and Generalized Unique Hard Attention (GUHAT) Transformers can only recognize the formal language of AC complexity class, while Averaged Hard Attention (AHAT) can recognize the language that they cannot recognize. Finally, literature [5] investigates the performance and limitations of Transformer in processing strings and sequences by introducing different positional coding and attention mechanisms and provides a unified framework to coordinate related findings.

Building upon the insights from the aforementioned literature, this study delves further into the limitations of language models in natural language processing. Initially, it conducts a systematic analysis of the current leading models, such as Transformers, focusing on their varying capabilities in representing formal languages and the factors that influence these abilities. By contrasting the impact of different model components—such as intermediate steps, attention mechanisms, and feed-forward networks—the researchers uncover potential bottlenecks in models' handling of complex linguistic phenomena. Subsequently, a series of training corpora constrained by syntactic motivations is constructed to measure the capabilities of distributed language models in distinguishing logical symbols. Experimental findings reveal significant difficulties in differentiating between logically distinct symbols, further substantiating the limitations of language models in comprehending deep linguistic semantics. Moreover, the study investigates whether neural network models can acquire the ability to recombine known basic words into more complex content while learning from data. Through comparisons of experimental results under various training strategies and model architectures, the researchers find that existing models still exhibit certain inadequacies in combining complex content. In conclusion, the limitations of current language models in handling formal languages and understanding linguistic meanings are summarized, along with challenges and prospects for future research. The study underscores the need for continued exploration of the intrinsic mechanisms and optimization methods of language models, as well as enhanced interdisciplinary collaboration and exchange, to jointly advance the technological progress and theoretical innovation in the field of NLP. This translation maintains the

DOI: 10.54254/2755-2721/97/20241301

academic tone and structure of the original Chinese text, ensuring a comprehensive and precise translation for academic or technical contexts. It accurately conveys the complex ideas and implications of the research while adhering closely to the original content.

2. Analysis of related research

Next, the paper will provide an overall explanation of the research results of several different papers, in order to more intuitively demonstrate the limitations of current LMs. To realize this goal, the paper starts with the capabilities already possessed by current LM such as Transformer [6], and summarizes them from two perspectives: known capabilities and limitations.

2.1. Differences in the expressivity of LMs

Literature [1] conducted a comprehensive study and investigation on the theoretical properties of Transformer in NLP. By constructing a unified framework, it systematically analyzed the differences in abilities of its different variants in expressing formal languages [1]. The literature is mainly based on the theory of formal language and regards Transformer as a string receiver or generator, where the inputs or outputs are treated as sequences of discrete symbols from a finite alphabet [1]. In the whole analysis process, the literature focuses on exploring the impact of Intermediate Steps, Attention [7], Feed forward Networks, Layer Normalization, Uniformity and Precision on the expressivity of Transformer and its variants.

Literature [1] suggests that there are three expressivities in the current prospects of NLP, namely decoders or encoder decoders with intermediate steps, encoders with average hard or softmax attention, and encoders with left hard or rightmost hard attention [1]. In the research, it was found that when it comes to the expressivity of transformer encoders, circuit complexity and logic are especially promising frameworks. [1]; Leftmost-hard-attention transformer encoders are in AC 0 and can't solve some intuitively simple problems, like Parity and Majority [1]; Softmax and average-hard attention give transformer encoders the ability to count. But they cannot solve problems like evaluating closed Boolean formulas. [1].

Based on the research of literature [1], the paper believes that appropriate training of models such as transformers is still the key to enhancing their attributes and expressivity. However, regarding the contradictory phenomenon of difficulty levels in handling related problems between models and humans, researchers should realize that there are fundamental differences in the learning logic process between models and humans. It is worth exploring whether the underlying learning logic based on the formation of human thinking is truly applicable to LMs and can help LMs make breakthrough progress.

2.2. Differentiation of Meaning

The previous research suggests that LMs in NLP only receive formal language training and therefore cannot understand the meaning of language [8]. Due to the belief that the form of language is often correlated with meaning [2], literature [2] further explores whether LMs can differentiate the meaning of language when they can only learn formal languages. Literature [2] focuses on propositional logic and generates various training corpus with different constraints to explore the ability of LMs to differentiate logical symbols ($\neg$, $\wedge$, $\vee$) under different constraints (semantic transparency), and to verify whether LMs can fully differentiate meanings under different constraints. Research has found that different training corpus attributes bring different performances and abilities to LM. However, no model built on a training corpus can clearly differentiate symbols with different meanings (the most semantically transparent training data did not enable models to separate the representations of symbols with similar forms but different meanings [2]). This clearly indicates that the current LMs have weaknesses in utilizing different semantic signals. Meanwhile, the literature [2] also provides a probable direction for future work.

The paper believes that this is a better extension of the viewpoint presented in 2.1. The current model is difficult to differentiate meanings, which has not been successful for the original intention of humans to establish models. However, the meaning of human language actually has no impact on the operation

DOI: 10.54254/2755-2721/97/20241301

of LM itself. In other words, LM has successfully completed language processing tasks without meaning, but the results may seem inappropriate when fed back to humans. Humans need meaning, but models don't. Therefore, if researchers expect to realize breakthroughs in the ability to differentiate model meanings, it is more appropriate to consider "meaning" as an external component of model functionality rather than an internal program operation. However, regarding the differences between models and humans, the paper also looks forward to further understanding: whether the differentiation of meaning by models should based on the meaning understood by humans, or it is possible to design a "meaning corpus" dedicated to models.

2.3. Compositionality

For LMs, its work needs to be close to human natural language, and the compositionality [9] of natural language is a noteworthy attribute. Literature [3] studied whether neural network models can acquire the ability to recombine known basic words into more complex content when learning data. From the perspective of formal language, the literature has generated a large number of datasets with controllable combinatorial properties using deterministic finite state transducers, exploring which attributes contribute to neural network learning of combinatorial relationships. By randomly sampling SFSTs (a restricted class of general finite-state transducers [10]) and sampling input-output pairs from each SFST, a unique string-to-string transformation dataset is created, and further investigating whether neural sequence-to-sequence models can learn Montague style combinatorial string-to-string transformations, where compositional behavior is defined to be homomorphic [3].

The research indicates that neural networks either tend to generalize completely or fail miserably, with almost no possibility of compromise. Meanwhile, the literature has established a simple complexity metric (transition coverage [3]) that can roughly forecast the learnability of SFST from sampled datasets.

The results of the literature seem to suggest that compositionality can also be transferred to the learning in language meaning of models, but according to the literature, the success of experimental learnability may also originate from the relatively limited synchronous context-free grammar [3], that is, the limitations of learning still exist to some extent, which requires further research by subsequent researchers.

2.4. Transformer Syntactic Challenges

The study [4] of the Transformer model replacing recurrent models in NLP tasks and its differences in modeling syntactic properties are investigated. By systematically studying the modeling abilities of Transformer in formal languages (especially counting languages and regular languages) and the role of its components in this process, it is found that Transformer performs well on certain subclasses of counting languages but has limited capabilities in regular languages. The role of the self-attention mechanism in modeling specific behaviors and the impact of positional encoding schemes on model learning and generalization capabilities are also discussed. Transformer performs well on certain subclasses of counting languages, but its capabilities in modeling regular languages are limited. It is found that compared to LSTM, Transformer performs poorly in dealing with languages requiring modeling of periodicity and module counting. In addition, positional encoding schemes have a significant impact on model learning and generalization capabilities. Transformer can generalize on certain counting languages, but its capabilities in modeling more complex regular languages are limited. This indicates that Transformer may also have similar limitations in dealing with natural languages.

The research results provide a new perspective on understanding the capabilities of Transformers in sequence modeling tasks and pose new questions for future research.

2.5. Transformer Language Recognition Limits

The research [5] analyzed three forms of self-attention mechanisms in Transformer encoders: UHAT, GUHAT, and AHAT. UHAT and GUHAT Transformers, as string acceptors, can only recognize formal languages in the AC0 complexity class, i.e., languages identifiable by families of Boolean circuits with constant depth and polynomial size. In contrast, non-AC0 languages such as MAJORITY and DYCK-

DOI: 10.54254/2755-2721/97/20241301

1 can be recognized by AHAT networks, indicating that AHAT is capable of recognizing languages that UHAT and GUHAT cannot. Moreover, it is shown that every UHAT can be simulated by an AHAT, establishing UHAT as a subclass of AHAT. This paper defined formal language recognition by Transformer encoders and demonstrated that the languages in UHAT and GUHAT can only be recognized by families of Boolean circuits with constant depth and polynomial size.

The primary contribution of the study lies in proving that GUHAT and UHAT can only recognize formal languages in the AC0 class, while AHAT is able to recognize languages outside the AC0 class. More formally, any language recognized using GUHAT can also be recognized by families of Boolean circuits with constant depth and polynomial size, establishing AC0 as the upper limit for the expressive power of UHAT and GUHAT. Additionally, the theoretical implications of these models in formal language recognition are discussed, and a general framework is proposed for comparing the three hard-attention Transformer models. Through theoretical analysis, limitations in the expressive power of Transformer models are revealed, offering new insights into the relationship between Transformer models and traditional computational models.

This translation closely adheres to the original text, maintaining its academic tone and technical detail. It provides an accurate and comprehensive translation suitable for academic or technical contexts.

3. Challenges and Prospects

Obviously, the current LMs for NLP are still in the stage of processing formal language tasks, and there are significant limitations in dealing with different aspects such as formal language and linguistic meaning [5], which has led to some problems in practical applications. Therefore, the limitations of LMs constantly remind researchers that there are still some fundamental contradictions in current natural language processing that need to be resolved. The paper encourages researchers to continue exploring the intrinsic mechanisms and optimization methods of LMs, while strengthening interdisciplinary cooperation and communication to jointly promote technological progress and theoretical innovation in the field of NLP. The paper also looks forward to future researchers making new progress in addressing the limitations of LM or proposing different fruitful opinions.

4. Conclusions

In the summary and analysis of the above literature, the paper investigates the limitations of LMs in NLP. Based on the above research results, mainstream models have both differences in the expressivity of formal language and significant difficulties in differentiating language meanings under different attribute training. Specifically, in the study of Transformer models, there are also many limitations that are difficult to overcome... These results truly demonstrate the limitations of current LMs in understanding deep semantics of language. Furthermore, after synthesizing the research results from multiple sources, the paper found that most of the limitations are not only reflected in the insufficient processing ability of the model itself, but also in the subjective needs of humans for the processing results of the model that can’t be met. For the former, this is an inevitable development experience of NLP, while for the latter, it is generated by fundamental differences between humans and models. In brief, models do not need to understand the results humans want from programs. The limitation of the latter is the difference between humans and models for understanding meaning and understanding code programs. And this limitation will surely affect the application effectiveness of current LMs. Therefore, the paper believes that increasing the level of attention to the differences between humans and models may be the key to breakthrough progress in natural language processing, or researchers may need to train language models based on more essential language attribute features.

Author Contribution

All the authors contributed equally and their names were listed in alphabetical order.

DOI: 10.54254/2755-2721/97/20241301

References

[1] Lena, S., William, M., Gail, W., David, C., Dana, A. (2024). What Formal Languages Can Transformers Express? A Survey. *Transactions of the Association for Computational Linguistics* 2024; 12 543–561. doi: https://doi.org/10.1162/tacl_a_00663.

[2] Traylor, A., Feiman, R., & Pavlick, E. (2021, August). AND does not mean OR: Using formal languages to study language models' representations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 158-167).

[3] Valvoda, J., Saphra, N., Rawski, J., Williams, A., & Cotterell, R. (2022). Benchmarking compositionality with formal languages. arxiv preprint arxiv:2208.08195.

[4] Bhattamishra, S., Ahuja, K., & Goyal, N. (2020). On the ability and limitations of transformers to recognize formal languages. arxiv preprint arxiv:2009.11264.

[5] Hao, Y., Angluin, D., & Frank, R. (2022). Formal language recognition by hard attention transformers: Perspectives from circuit complexity. Transactions of the Association for Computational Linguistics, 10, 800-810.

[6] Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N., et al., (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS).

[7] Dzmitry, B., Kyunghyun, C., & Yoshua, B. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR.

[8] Emily, M., and Alexander, K. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198. Association for Computational Linguistics.

[9] Richard, M. (1970). Universal grammar. Theoria.

[10] Mehryar, M. (1997). Finite-state transducers in language and speech processing. Computational Linguistics, 23(2):269–311.yuan8