Book Review

Cognitive Plausibility in Natural Language Processing

Lisa Beinborn and Nora Hollenstein
(Vrije Universiteit Amsterdam & University of Copenhagen)

Springer (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst), 2024, xi+158 pp; hardback, ISBN 978-3-031-43259-0; ebook, ISBN 978-3-031-43260-6; doi:10.1007/978-3-031-43260-6

Reviewed by
Yevgen Matusevych
University of Groningen

Recent successes in Natural Language Processing (NLP) give rise to more and more computational models aimed at generating and understanding human language. Traditionally, researchers evaluate such models by looking at how accurate they are at a given task, but the focus in the field is slowly shifting to evaluating other characteristics, such as models' fairness, interpretability, efficiency, and so forth. An often overlooked aspect is a model's **cognitive plausibility**: that is, how "human-like" a model is. Cognitive plausibility is a multifaceted concept grounded in the field of cognitive science, which often uses computational models to study various aspects of human cognition, including language production and comprehension. The present book is a contribution to narrowing the wide gap between state-of-the-art NLP architectures and human cognition.

In the quickly changing world of NLP, it is not easy to summarize recent advances, nonetheless, this book's comprehensive selection of examples from various studies offers a solid overview of the current research on models' cognitively plausibility. The examples are combined with discussions of theoretical and methodological issues at the interface of NLP and cognitive science. Moreover, each of the main content chapters ends with an overview of relevant ethical issues, a necessary consideration in the era of powerful language models. The book will be compelling reading to NLP researchers interested in human cognition and model interpretability, as well as to cognitive scientists and psycholinguists willing to better understand computational modeling approaches in the language domain. Gradual presentation of the material, with two introductory chapters (see below), also makes the book generally accessible to students and those with only basic knowledge of NLP.

The book consists of seven chapters. Chapter 1, "Introduction," explains why cognitive plausibility is important to consider in NLP: In addition to obvious advantages of cognitively plausible models for cognitive science, the behavior of such models is more intuitive to interpret for human speakers. Exploring this link between cognitive plausibility and interpretability is one of the book's goals, and it is repeatedly—and very

https://doi.org/10.1162/coli_r_00517

successfully—exploited throughout all the chapters. Another important message in the introduction is that cognitive plausibility is a graded concept that involves multiple dimensions, and this book focuses on three of them: the similarity between the decisions made by models and humans, between the representational structures they use, and between their procedural strategies.

The five chapters with the main content can be divided into two large sections: the more introductory ones (2–3) and more in-depth ones (4–6). Specifically, Chapter 2, "Foundations of Language Modeling," provides an overview of basic concepts in language modeling, such as conditional probability of a sentence, model perplexity, recurrent neural networks, pretrained language models, and so on. Particular emphasis is made on modeling choices that link language models to human language processing: For example, a model's objective function and architecture can determine whether the model processes words sequentially or not, while input units that models are trained on may or may not reflect the underlying linguistic structure. Chapter 3, "Cognitive Signals of Language Processing," introduces the types of data that can be used for evaluating cognitive plausibility of computational models. Here, the reader is made aware that the focus of the book is on language comprehension rather than production, which some may find slightly disappointing given that many state-of-the-art generative language models are celebrated largely for their ability to produce human-like language. On the brighter side, even in comprehension there is plenty of relevant data sets, which capture both human speakers' behavioral responses and their brain activity patterns. Moreover, one can combine different types of data for an even more comprehensive model evaluation. Importantly, many data sets collected from human speakers need to be preprocessed or otherwise adapted to NLP settings, and this chapter also offers an overview of common techniques in this area.

The following three chapters are more in-depth and discuss the three dimensions of cognitive plausibility mentioned above. Chapter 4, "Behavioral Patterns," largely focuses on model interpretability methods through the cognitive lens. First of all, one can analyze a model's input and output: properties of the input data, model's behavior on well-defined subpopulations of data and its performance on specific instances depending on their difficulty. Second, there is a variety of tests or even out-of-the-box test suites for targeted evaluation of NLP models. Many of them focus on models' linguistic abilities, while others (e.g., occlusion or perturbation tests) are designed to stress-test models' robustness on intricately designed examples. Here, the authors call for using finer-grained model evaluation and for developing multilingual models that are not optimized for English data, as is often the case in NLP. Another promising approach to narrowing the gap between models' and human speakers' behavior, according to the authors, consists in designing cognitively plausible curricula for model training, but sadly, in the 2023 BabyLM challenge curriculum learning methods only resulted in modest improvements (Warstadt et al. 2023), and the jury is still out on this subject.

While the previous chapter focuses on models' inputs and outputs, Chapter 5, "Representational Structure," discusses interpretability of models' *internal* representations. Neural representations are commonly expressed as vectors in a high-dimensional space, and measuring similarity between them is central to research in this area. From the cognitive perspective, however, similarity—even between words, let alone longer units—is a complicated concept: units can be judged similar for a variety of reasons, which also highly depend on the context. Moving beyond a single representational space, one can also measure the similarity of different spaces, including the ones derived from human speakers' behavioral or brain responses. In many cases, it is crucial to have a concrete hypothesis about a model's representations and test it in a targeted way—for example, using probing classifiers, a common method for finding out whether a specific feature is encoded in a model's representation space. Another fruitful research direction is mapping models' representations to brain responses: For example, can a probing classifier learn to predict brain activation patterns?

The final dimension of cognitive plausibility is discussed in Chapter 6, "Procedural Strategies." Strategies that a model adopts are determined by its architecture. For the time being, transformers are by far the most common architecture in NLP, and a large part of this chapter discusses the mechanisms of attention and self-attention used in transformers. Somewhat sadly from this book's perspective, multi-head attention can hardly be considered cognitively plausible, but nevertheless, studying attention weights can still help us understand relative importance of input units (e.g., words), and the model's importance values can be compared to human data, such as gaze patterns during reading. Overall, this chapter mostly focuses on sentence processing tasks, since they provide a fruitful ground for studying various procedural strategies and effects: incremental processing (cf. Chapter 2), priming, hierarchical processing, and so on. Concrete proposals to improve the cognitive plausibility of models' algorithms include the use of multi-task and transfer learning setups, through the explicit integration of human problem-solving strategies into models' training process.

The final Chapter 7, "Towards Cognitively More Plausible Models," is a brief recap of the book. Although the authors conclude they could not find a silver bullet to make a model cognitively plausible on all the three dimensions considered, they nevertheless successfully propose concrete methods for designing more cognitively plausible models. Among others, these proposals include taking into account instance difficulty in model evaluation, developing more context-aware tools for representation analysis, integrating information from multiple modalities into the models, and adopting a truly multilingual perspective on model design.

One key takeaway from this book is that cognitive plausibility is a complex concept—not only because there are several dimensions to it, but also because data collected from human speakers is often less straightforward than what's dictated by existing NLP models' objective functions. The authors provide plenty of examples of this complexity throughout the book: Human speakers can disagree in their linguistic annotations, their conceptual representations tend to be fluid, their similarity judgments are nuanced and graded. One possible way forward for NLP is to embrace this uncertainty of human language behavior, an avenue that the field is only starting to explore (e.g., Baan et al. 2023; Liu et al. 2023).

References

Baan, Joris, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807.

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.

Yevgen Matusevych is an assistant professor at the Center for Language and Cognition Groningen (CLCG), University of Groningen, the Netherlands. He has worked on computational cognitive models of language learning, in particular bilingual and non-native learning, across multiple linguistic domains, from speech perception and lexical-semantic organization to morphology and grammar. His e-mail address is `yevgen.matusevych@rug.nl`.