Synthesis Lectures on
Human Language Technologies

SYNTHESIS
COLLECTION OF TECHNOLOGY

Lisa Beinborn
Nora Hollenstein

Cognitive Plausibility in Natural Language Processing

Springer

Synthesis Lectures on Human Language Technologies

The series publishes topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Lisa Beinborn · Nora Hollenstein

Cognitive Plausibility in Natural Language Processing

Springer

Lisa Beinborn
Faculty of Humanities
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Nora Hollenstein
Department of Nordic Studies and Linguistics
University of Copenhagen
Copenhagen, Denmark

ISSN 1947-4040 ISSN 1947-4059 (electronic)
Synthesis Lectures on Human Language Technologies
ISBN 978-3-031-43259-0 ISBN 978-3-031-43260-6 (eBook)
https://doi.org/10.1007/978-3-031-43260-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Acknowledgments

The development of this book spans multiple years of research. We extend our heartfelt gratitude to our collaborators and colleagues for all the inspiring discussions, for pointing us to new ideas, and for patiently contradicting us. Our writing journey began after teaching a summer course together with Willem Zuidema at the European Summer School for Logic, Language, and Information in 2021. We thank Willem for sharing his visionary ideas and for organizing the plenary discussion with the enthusiastic students that inspired this book.

Research is always a collaborative endeavor and our colleagues serve as an invaluable sounding board for shaping our thoughts. To name only a few collaborators, we would like to highlight the incredibly talented Ph.Ds. Charlotte Pouw and Ece Takmaz and our excellent co-authors Lena Jäger, Maria Barrett, and Stefanie Brandl. Your perspectives helped us better understand the many facets of cognitive plausibility.

Lisa deeply appreciates the privilege of being embedded in Amsterdam's vibrant NLP community. She expresses her gratitude to the researchers at the CLTL group at VU Amsterdam, and at the CLC and dialog modeling groups at the ILLC whose brilliance and dedication have been a constant source of inspiration. And she thanks her family for their patience because they matter most.

Nora is grateful for the stimulating working environment at the University of Copenhagen. She is especially thankful to her colleagues at the Centre for Language Technology, to Anders Søgaard for his guidance, and to the students of the MSc IT & Cognition program. She thanks the researchers at the Department of Computational Linguistics of the University of Zurich, a continuous pillar of support from the very beginning of her NLP journey. And, always, her husband Simon.

Graeme Hirst provided patient and motivating guidance through the process of writing our first book. We thank him and the anonymous reviewers who gave us extraordinarily constructive and encouraging feedback with great attention to detail.

Lastly, we genuinely hope that you enjoy reading this book as much as we enjoyed our collaboration over the past few years.

July 2023

Lisa Beinborn
Nora Hollenstein

Contents

About the Authors

Lisa Beinborn is an assistant professor for Natural Language Processing at the Computational Linguistics and Text Mining Lab at Vrije Universiteit Amsterdam. Her research focuses on cognitively plausible language processing and cross-lingual models. She studied Computational Linguistics in Saarbrücken, Barcelona, and Bolzano, and obtained her Ph.D. in Computer Science at TU Darmstadt. She works on the interpretability of representational transfer and is interested in educational applications of NLP.

Nora Hollenstein is currently working at the Center for Language Technology of the University of Copenhagen and at the Department of Computational Linguistics of the University of Zurich. She obtained her Ph.D. from ETH Zurich working on cognitively inspired NLP. The focus of her research lies in improving and evaluating natural language processing applications with cognitive signals such as eye-tracking and brain activity recordings. She is especially interested in multilingual and multimodal NLP.

Introduction

1

Language is a powerful tool of human communication that provides elegant mechanisms for expressing highly complex phenomena. We use language every day and in all aspects of our lives. The versatility and variability of language make it a difficult subject for computational modeling, in contrast to more systematic sensor signals. Language follows underlying rules only to surprise us with exceptions and ambiguities on all linguistic levels and understanding its subtleties requires even more culture-specific knowledge than interpreting images.

We cannot derive an accurate static description of language because it dynamically evolves over time and across domains. More than 7,000 signed and spoken languages exist in the world covering a large spectrum of typological configurations [1, 2]. If we try to isolate a fundamental principle of language processing in scientific models, we will soon encounter a language with a complementary linguistic structure that creatively contradicts our assumptions.

In spite of these complexities, humans usually process language effortlessly. We are able to vary our language use to smoothly adapt to the target audience and dynamically integrate situational cues for seamless disambiguation. Natural language processing (NLP) research has already spent decades trying to understand how to computationally model language but complex reasoning tasks and creative constructions still lead to obvious failures of models. Nevertheless, the success of the field is undeniable. It attracts a continuously growing number of researchers and language processing models have become a key technology in our daily lives. These developments are strongly linked to the increasing availability of large amounts of training data and more efficient computing resources. Neural language models (LMs) are trained on terabytes of data and optimize millions of parameters to extract patterns from text.

When we train such a model to represent language, we fall back on a range of simplifying assumptions. For text-based models, we often expect that the text is formatted as one sen-

L. Beinborn and N. Hollenstein, *Cognitive Plausibility in Natural Language Processing*, Synthesis Lectures on Human Language Technologies,
https://doi.org/10.1007/978-3-031-43260-6_1

tence per line, does not contain any images or special fonts, and is (mostly) typo-free. We implicitly assume that large text corpora are representative of modern language use and that optimizing the implemented language modeling objective relies on information that is relevant to understanding language. As researchers, we are usually aware that our assumptions oversimplify realistic scenarios. When we develop computational models, these assumptions become explicit in our modeling choices which makes it possible to directly object to them and empirically compare competing hypotheses. Building on the famous aphorism by Box [3] stating that "all models are wrong", Smaldino [4] discusses how "stupid models" can provide important insights into our misconceptions. He claims that working with computational models forces us to expose "our foolishness to the world" as a "price of seeking knowledge".

While the general benefit of computational modeling has been clearly established in multiple disciplines, we observe substantial disagreement in evaluating which properties characterize a "useful" model. Ruder et al. [5] describe how NLP research is only slowly evolving from purely performance-oriented experiments to integrating additional factors such as fairness, interpretability, computational efficiency, and multilingualism. Throughout this book, we propose integrating cognitive plausibility as an additional factor and agree with their call for more multi-dimensional research to capture interactions. We think that a multilingual perspective is required to learn more about cognitively plausible principles of language processing. And cognitively more plausible models can lead to computationally more efficient models.

Cognitive plausibility itself is a multi-faceted concept that varies considerably across disciplines. Computer scientists focus on the quantitative performance of the model, which should not be distinguishable from humans on static benchmark datasets. Neuroscientists focus on the biological plausibility of the model and aim to develop models of synaptic plasticity, which are evaluated on toy datasets that are much smaller than the common evaluation datasets in natural language processing. Psychologists focus on the plausibility of the learning processes in language models. They question the size and quality of the input data, examine memory and attention constraints, and explore learning curves. They try to isolate experimental factors by working with carefully designed stimuli that are often not representative of realistic language use.

We approach the concept of cognitive plausibility by diving into interpretability research and identifying the potential of its methods for cognitively inspired research questions. We focus on computational language models that are based on neural network architectures. These models are very powerful and their distributed approach is often motivated as being more cognitively plausible. Unfortunately, the expressivity of the model is gained at a loss of transparency. The high-dimensional matrix transformations that characterize neural models are hard to conceptualize for humans. When comparing models with millions of parameters, it becomes difficult to isolate the underlying modeling assumption that explains the differences. Interpretability methods are being developed to gain a better understanding of the inner workings and the inductive biases of neural models. These methods are often dis-

cussed from an engineering perspective focusing on quantifiable and measurable approaches to compare different models. We take a different stance: we explore interpretability methods through a cognitive lens by linking them to aspects of human language processing.

In this chapter, we first discuss why cognitive plausibility is a relevant factor for natural language processing research. We then define three dimensions of cognitive plausibility that we address in this book and provide an overview of how we approach their analysis.

1.1 Does Cognitive Plausibility Matter?

The cognitive plausibility of a model is an aspect that is commonly only implicitly targeted in natural language processing by assuming that higher average performance is obtained by better models and that better models are cognitively more plausible. We think that it is worthwhile to address cognitive plausibility more explicitly and want to initiate a more nuanced and in-depth discussion.

1.1.1 Human-Centered Natural Language Processing

In a recent data-driven survey on the values encoded in highly cited machine learning publications, the goals of developing a "human-like mechanism", "learning from humans" and being "transparent (to users)" can be found at the bottom of the list, while performance is the main driving force in 96% of the examined papers [6]. NLP research is strongly influenced by trends and developments in machine learning research but humans are central to our field: language is generated and developed by humans and language processing tools are directly used by humans (in contrast to neural models that are used for building machines or cultivating plants). Both, the input and the output of our models are characterized by human preferences, thus "human language processing is the ultimate gold standard for computational linguistics" [7].

Ethayarajh and Jurafsky [8] criticize that the progress in computational modeling is currently determined mostly by performance-oriented comparisons and propose to develop more user-centric leaderboards. They take a micro-economic approach to identify a model's utility for an end user and urge for more transparent reporting of practical factors such as model size, energy efficiency, and inference latency. In our opinion, the cognitive plausibility of a model is an underestimated factor of its utility.

The cognitive and economic aspects of model utility are not as unrelated as they may seem. Humans still outperform language models with respect to their linguistic generalization capabilities. We learn, understand, and produce language effortlessly, and while the cognitive mechanisms are not yet fully understood, it is clear that we do so very efficiently. As current computational models of language still struggle with phenomena that pose no problems for humans, it seems an obvious choice for us to turn to psycholinguistics and neurolinguistics

for inspiration to build better computational models. If we want our models to acquire language as efficiently as humans and to generalize to new structures and contexts [9], we should reward cognitively more plausible architectures that simulate human transfer skills rather than models that excel in capturing statistical patterns of limited datasets [10].

1.1.2 Understanding Language Versus Building Tools

Natural language processing is an interdisciplinary research area by definition. It attracts researchers from a wide range of diverse backgrounds with very different research ambitions. When we develop computational models of language, two main goals can be distinguished. We might be driven by the vision to develop a computational model to better understand how humans process language. If we take a more practical perspective, we aspire to build a tool that automates language-related tasks to simplify our daily routines. While cognitive plausibility is clearly central to the first goal, it is less obvious for the tool-oriented approach.

One could argue that the cognitive plausibility of a model is irrelevant if the model works well enough for an intended use case. An information retrieval model that uses static keywords and templates can extract valuable information from scientific papers. While such a model can be useful for the average scenario, the user needs to take conscious countermeasures to account for its weaknesses (i.e., checking for alternative keywords and ascertaining that the extracted information is not framed within a negative context).

When working with neural language models, it becomes challenging to identify the patterns that determine the behavior of the model. Language input is complex and highly ambiguous and the model's decisions are learned by optimizing millions of parameters. With these two factors combined, it is almost impossible to compile clear descriptions for users in the form: if the input has property *X*, then the expected outcome quality is *Y*. The addition or omission of a single word might already flip the expected output label, but the model might not be sensitive to such changes if they have not been seen in the training data.

We are convinced that the ability to robustly anticipate the strengths and weaknesses of a model is a decisive factor in human-centered NLP. Cognitively more plausible models exhibit decision patterns that are more intuitive for human users because they are consistent with what they would expect from a colleague. We assume that cognitive consistency facilitates the practical application of a model because users can anticipate the reliability of its predictions and keep an eye on potential sources of error. For example, an essay scoring system that assigns a high grade to a well-written essay although it contains an argumentative fallacy is more plausible than a system that rewards an essay that is simply a bag of keywords. A tired human might have made the same mistake in the first case but could easily spot the word salad.

Language models have become the Swiss army knife of natural language processing because the finetuning methodology enables versatile adaptation to many tasks. High performance on a challenging subset of these tasks is often claimed to be an indicator of natural

language understanding. Bender and Koller [11] claim that understanding requires a representation of meaning and initiated a discussion on whether meaning can be derived from form alone. They argue that meaning can only be interpreted with respect to the communicative intent of the utterance and define the conventional meaning of an expression as an abstraction over all possible contexts. As language models do not have access to functions of consciousness and self-awareness, they cannot capture forms of communicative intent. They can derive patterns from the training data but they cannot infer commonsense knowledge that is not explicitly expressed in textual sources due to the reporting bias [12]. The development of language models is advancing at a very fast pace and newer models master tasks that have been considered impossible to achieve. They can process and generate language to an extent that indicates a broad and growing set of linguistic competencies but we cannot conclude that they understand the meaning of words since they still fall short in many other respects such as grounding meaning in perception and action [13].

Our book presents a structured introduction to methods for analyzing the cognitive plausibility of language models. We do not aim to provide an absolute definition of the upper bound for a cognitively plausible model because we think that it is more useful to view cognitive plausibility as a graded concept.

1.1.3 Ethical Considerations

The release of ChatGPT [14] has led to a sudden recognition of language models reaching far outside academic target groups. With the wave of hastily released premature applications building on proprietary technology, the urgency of ethical scrutiny has become undeniable. Natural language processing is particularly sensitive to ethical problems because the way we use language to frame events, opinions, or feelings, affects our moral judgments and our perception of responsibility [15]. When language models are becoming increasingly cognitively plausible, we need to assure that they are not used in harmful ways. Ethical aspects should be considered at all levels of model implementation, starting with the data up to the misinterpretation of model outputs and the potential misuse of applications. In each chapter of this book, we discuss ethical aspects related to the respective methodology.

We discuss the problems that arise due to biases in the training data of language models and point out sustainability concerns with respect to computationally expensive training regimes. Many of the methods presented in this book rely on cognitive signals collected from human participants. When dealing with such sensitive data, it is imperative to adhere to privacy regulations. We discuss aspects of anonymization and overgeneralization and draw attention to systematic demographic biases. We expand on the problem of societal biases and discuss the trade-off between normative and descriptive ethics with respect to model behavior. We address the need for transparency about the limitations of our methodological choices and the importance of communication and traceability to ensure open pathways between academic research, society, and education.