The Role of Explanatory Value in Natural Language Processing

Kees van Deemter.
Utrecht University

A key aim of science is explanation, yet the idea of explaining language phenomena has taken a backseat in mainstream Natural Language Processing (NLP) and many other areas of Artificial Intelligence. I argue that explanation of linguistic behaviour should be a main goal of NLP, and that this is not the same as making NLP models “explainable”. To illustrate these ideas, some recent models of human language production are compared with each other. I conclude by asking what it would mean for NLP research and institutional policies if our community took explanatory value seriously, while heeding some possible pitfalls.

1. Introduction

In this short paper, I will argue that much recent work in Natural Language Processing (NLP) has focussed too narrowly on the performance of its models, as measured by various intrinsic or extrinsic evaluations, neglecting some vitally important other dimensions of model quality, which I will collect loosely under the header “explanatory value”. Before explaining what I mean, let me set the scene.

Broadly, NLP can be pursued in three mutually overlapping ways, which emphasize different aspects of our work. First, there is NLP as Engineering, where NLP models are built primarily to serve some practical goal, answering a need that exists in society. Second, there is what might be called NLP-as-Mathematics, which studies algorithms and models in their own right, comparing them and developing new ones. Finally, there is NLP-as-Science, where models are constructed with the aim of expressing, testing, and ultimately enhancing humankind’s grasp of human language and language use, because computational models offer a level of explicitness and detail that other theories of language often lack. For example, a sentiment analysis model may be seen as a highly explicit theory of the ways in which psychological states manifest themselves in language; a summarisation model can be seen as a theory of what is most informative in a text; a Machine Translation model may be seen as a theory of translation.

I will focus primarily on NLP-as-Science arguing that, if explanation is our aim, then this should include several dimensions. In particular, if we rely solely on the performance of our models, then we risk building models that are ad hoc, that are unwieldy, that are difficult to link with existing insights, and that do not allow us to answer counterfactual questions such as, “How well would this model perform if we applied it to texts of a different genre?”

One might argue that a lot of work in NLP focusses on explanation already, because it aims for explainability (e.g., (Ghassemi, Oakden-Rayner, and Beam 2021)). This argument, however, confuses (1) explaining a natural phenomenon (e.g., an aspect of language use) and (2) explaining a model (i.e., a piece of software). Explanation is about

---

1 Although I am conscious that this use of the term “explanatory” is broader than its daily usage, I will argue in section 2 that it makes sense to use the term in this way.
"What principles underlie this phenomenon?", whereas explainability is about "Why did this model make these decisions?" The difference is starkest when the model doesn’t match the phenomena very well. Suppose you had a model for classifying student essays into good (pass) and bad (fail). Suppose the model had terrible performance but excellent explainability. The model would thus be highly explainable, and this could be useful for a stakeholder who wonders whether to trust its decisions, or a developer wanting to improve it. Yet these explanations would not tell us what makes an essay good or bad because (we assumed) the model does not know the difference. Similar things can be said about the idea that model evaluation should include a systematic analysis of errors (e.g., (McKeown 2020); (Ribeiro et al. 2020)). The importance of finding out what a model gets right and wrong can hardly be overstated but, by itself, it can only shed light on the model, not the language phenomena we are seeking to understand.

2. Dimensions of Explanatory Value

Explanation plays a key role throughout the sciences (Popper 1934; Overton 2013; Woodward and Ross 2021) and in daily life (Lombrozo 2006). Accordingly, many disciplines have seen lively discussion of what it means for a theory or model to offer a good explanation of data, but in recent NLP there has been little discussion about such matters. Consequently, it is unclear what makes a good explanation, and whether explanation even matters in NLP. Based on a liberal borrowing from other disciplines, I will discuss what different dimensions explanation can involve, what these might mean for assessing NLP models, and what the implications would be if our community took them seriously.

**Performance.** One dimension of explanatory value is *performance*, which includes metrics such as Precision, Recall, DICE, BLEU, Moverscore (Sai, Mohankumar, and Khapra 2022), (Celikyilmaz, Clark, and Gao 2020), which allow researchers to compare a model’s predictions with a gold standard. Evaluations based on human judgments (e.g., (van der Lee et al. 2019)) or task-based evaluations are varieties of performance likewise. Performance is naturally thought of as a component of explanatory value. For instance, if model A has better performance than B, then other things being equal, A has greater explanatory value than B. If a model does not allow performance to be assessed at all, then I will consider such models to have low performance.

A natural complement to performance is a group of factors I will call “support”. Simply put, when a model is tested in an experiment, support is any evidence for the model other than the evidence from that particular experiment. I will distinguish between (what I will call) direct and indirect support.

**Direct support.** Direct support assesses a model’s ability to make accurate predictions for unseen data, and to generalize to related tasks and different contexts. It comes from a plurality of broadly similar experiments. Suppose a caption generation model was tested on a set of holiday snaps, showing good performance. Direct support could

---

2 For instance, via computer-generated “rationales” that highlight text fragments that were particularly important for each classification decision, e.g. Lei, Barzilay, and Jaakola (2016).

3 This includes repetitions, reproductions, and replications, using the terminology of the ACM at www.acm.org/publications/policies/artifact-review-and-badging-current.

4 Caption generation models are Vision and Language models that generate textual captions for visual images (Hodosh, Young, and Hockenmaier 2013; Agrawal et al. 2019).
include an experiment on another collection of holiday snaps, an experiment with a different type of images, or a probe that investigates whether the success of the model is due to some accidental feature of the dataset. Similarly, if a Machine Translation model is tested on a new language pair, then this can be seen as offering direct support to the model. Direct support views language corpora as data samples that are only of interest to the extent that they are representative of a wider population of data. This implies that we should ask ourselves what type of language use we want our corpora to be representative of, and be on our guard against confounding variables (i.e., accidental circumstances that may have affected our results Ribeiro, Singh, and Guestrin (2016)).

**Indirect support.** A dominant theme in the Philosophy of Science is that explanation should involve a reduction of the unknown (e.g., some previously unobserved facts) to something already known, such as an existing law or insight or model (Hempel and Oppenheim 1965). These ideas lie at the heart of the scientific enterprise (see e.g. Hepburn and Andersen (2021)), understood as “the attempt to understand the world around us” (Levesque (2014), speaking about Artificial Intelligence).

For example, suppose a study finds that a medicine $X$ damages patients’ liver. Suppose we also know $X$ contains a molecule $Y$ which is known to be toxic, then this existing insight offers indirect support to the finding about side effects. Likewise, if an aspect of language use can be shown to enhance or speed up communication, this can lend additional credence to a model that incorporates this aspect. Support can also be negative. For example, if a physics model explains certain observations by positing “action at a distance”, (i.e., where an object can be affected by another object without being “touched” by it in any way) then the difficulty of making sense of that idea has been seen as diminishing the value of the model (Berkovitz 2008).

Indirect support is essential to what is called explanation in everyday parlance, and can even give us a sense that we “understand” the underlying mechanism (for instance when we know that the above molecule $Y$ kills a particular liver enzyme), but the idea of mechanism has been notoriously difficult to underpin (Craver and Tabery 2019). Indirect support in NLP can sometimes take the form of cognitive plausibility (although it does not have to, see e.g. Harnad (1989)). Suppose, for example, a text comprehension model uses an algorithm that is NP-complete, then this makes it implausible as a model of human behaviour because it suggests a brain mechanism that would be so time consuming that it could not work in practice.

When indirect support $I$ is invoked, two key questions are in order: How certain are we about $I$ (e.g., what is the evidence or reasoning underlying $I$?), and To what extent would $I$ (if correct) support the model? Most scientists accept that higher principles have a role to play, though an assessments of whether a principle is rightfully invoked can be a matter of vigorous debate. For instance, action-at-a-distance has become an accepted part of physics despite being seen as implausible for a long time.

**Parsimony.** Parsimony is the idea that a simpler model is a better model. Parsimony is closely related to Ockham’s Razor and to the idea that models should be as elegant as possible (see e.g. Greene (2000)), which counts philosophers and physicists such as Karl Popper and Paul Dirac among its early proponents. Parsimony is an aspect of explanatory value because if we do not insist on some form of parsimony, a model could be deemed to be highly explanatory even if it was nothing more than a large collection of isolated facts or rules without any attempt at generalisation.

Regarding the question of how scientific elegance should defined, there are different views. In particular, parsimony can concern different aspects of the model (see e.g.
Fitzpatrick (2022), section 3). Accounts of parsimony that focus on the number of postulates employed by the model, for example, have been defended on the grounds that more parsimonious models have a greater probability of being true. A type of parsimony more relevant to NLP says that, even if two models cannot (or: not very clearly) be distinguished in terms of their performance, then if one is simpler than the other, the simpler model should be preferred. This type of parsimony is routinely used, and sometimes defended explicitly, by syntacticians (Brody 1995; Akmajian and Heny 1975), among others.5

Invoking parsimony can be risky, particularly if a complex model is dismissed that has better performance than its competitors. Physicists such as Sabine Hossenfelder believe elegance has played too large a role in discussions of string theory (Hossenfelder 2018). Nonetheless, the idea that a lack of parsimony can diminish the value of a theory is widely accepted. We will count it as a fourth dimension of explanatory value.

A Bayesian perspective on the progression of science (e.g. (Jaynes 2003)) may help to clarify these dimensions. Let $D$ be the data obtained when a model $M$ is tested, and $X$ is everything else we know, including indirect support for or against $M$. Then performance of a probabilistic model $M$ can be seen as $P(D|M,X)$, the probability $D$ would have if $M$ were true. What one is typically interested in is $P(M|D,X)$, the probability of $M$ given $D$ and $X$. Direct support is Bayesian update, where more and more data sets $D_1, \ldots, D_n$ are brought to bear, yielding $P(M|D_1, \ldots, D_n, X)$. Even parsimony can, at least in principle, be captured along Bayesian lines, by using Solomonoff’s Prior (Solomonoff 1964; Hutter, Legg, and Vitanyi 2007), which assesses the complexity of a model by measuring its a priori probability (i.e., the probability of the model before any data are considered).

3. Case study: Two types of Referring Expressions Generation

To illustrate both the usefulness and the pitfalls of assessing the explanatory value of NLP models, I examine two types of referring expressions generation (REG). I choose REG because referring is an essential part of human communication that has been studied from many different angles, using very different types of models; moreover, the performance of REG models has been tested extensively, and the outcomes of these tests will inform our discussion of the explanatory value of these models.

3.1 Generating one-shot Referring Expressions

One-shot REG has been much studied in NLP ((Dale 1989), (Dale and Reiter 1995), (Krahmer and van Deemter 2012), (Yu et al. 2016), (Luo and Shakhnarovich 2017)): the research question is, given a “scene” composed of objects, and without any linguistic context (hence “one-shot” REG), what properties do human speakers select when they refer to one of the objects in the scene? The patterns observed here are far from trivial, and sometimes counter-intuitive (van Deemter 2016). Here I concentrate on a class of models that emerged from controlled experiments involving simple artificial scenes whose objects have well-understood properties (shape, colour, size, etc.) that can be

---

5 The opening chapter of Akmajian and Heny (1975) uses this example: X liked you is assigned the underlying form X did like you, because this allows one to generate tag questions (X liked you, didn’t he?), negated sentences (X did not like you) and emphatic sentences (X did like you) using one and the same mechanism, thereby minimizing the complexity of the grammar.
manipulated precisely by the experimenter and presented to participants on a computer screen. Such experiments trade away some of the complexity of real-world scenes to allow a maximum of experimental control.

We compare five models. One model is an application (which I will call RSA-REG) of Frank and Goodman’s Rational Speech Act (RSA) model (Frank and Goodman 2016), (Frank and Goodman 2012). RSA is formalisation of the Gricean idea that communication is always optimally cooperative; consequently, RSA-REG’s speaker model emphasizes discriminatory power: the likelihood that a property is chosen for inclusion in a Referring Expression (RE) is proportional to its discriminatory power (i.e., the proportion of scene objects to which the property does not apply).

The other models grew out of a “Bounded Rationality” research tradition that emphasises the idea that is skeptical about the idea that speakers routinely compute discriminatory power when they refer. A well-known version of this experimentally well-supported idea (e.g. (Belke and Meyer 2002)) is the Incremental Algorithm of Dale and Reiter (1995), which assumes that properties are arranged in a linear sequence that lists them according to the degree to which they are preferred. A range of experimental findings Koolen et al. (2011); Gatt et al. (2013); van Deemter (2016); Van Gompel et al. (2019)) led to various improvements, including two probabilistic versions of the Incremental Algorithm, and our own model called Probabilistic Over-specification (PRO), which combines discriminatory power with a probabilistic use of preference.

Comparison 1: Performance. In (Van Gompel et al. 2019) we reported an experiment in which the PRO model outperformed the other algorithms in terms of the human-likeness of their output.

Comparison 2: Direct support. Algorithms in the Bounded Rationality tradition have often been tested, including the evaluation campaigns of Gatt and Belz (2010). Direct support for RSA-REG does not yet reach the level of the other models; naturally, support for these models may grow over time.

Comparison 3: Indirect support. At first sight, there is much indirect support for RSA, given the intuitive appeal of describing human behaviour as rational. However, a wealth of work in behavioural economics has shown that rational behaviour is affected by time and memory limitations, necessitating shortcuts (Elster 1983; Simon 1991; Gigerenzer and Selten 2002; Gershman, Horvitz, and Tenenbaum 2015), and other deviations from rational behaviour (Kahneman and Tversky 2013). Experiments on REG are in line with these findings (van Deemter 2016; Van Gompel et al. 2019). For example, PRO is full of shortcuts that avoid the arduous computation of the discriminatory power of each property that would be required by the RSA-REG algorithm.

Comparison 4: Parsimony. The computational core of RSA-REG can be written in just two simple equations; by contrast, PRO is a rule-based algorithm whose pseudo-code needs about a page. It seems reasonable, therefore, to say RSA-REG is more parsimonious than PRO.

---

6 The mechanisms of (Degen et al. 2020) could probably lend RSA-REG better performance (Rubio-Fernandez 2021), but until a systematic performance assessment of the resulting model is available, Frank and Goodman’s model will serve our illustrative purposes.
3.2 Generating Referring Expressions in Context

REG-in-Context is another well-studied area of REG. It focusses on co-reference in discourse. It often starts from a text in which all referring expressions (REs) have been blanked out; it predicts, for each of these blanks, what RE should fill it. Other than the identity of the referent, the main information for the model to consider is the sentences around the RE, because this guides the choice between pronouns, proper names, and descriptions. The other entities mentioned in the text play a role not dissimilar to the “distractor” objects displayed on a computer screen in One-shot REG (previous section).

A long tradition of linguistic research has led to theories such as accessibility theory (Ariel 1990), the givenness hierarchy (Gundel, Hedberg, and Zacharski 1993), and Centering Theory (Brennan 1995). These theories emphasise the effect of the recency of the antecedent (e.g., in terms of the number of intervening words), its animacy (animate/non-animate), and the syntactic structure of the sentences (e.g., Does the RE occur in the same syntactic position as the antecedent?) Computational accounts can be classified in terms of whether they use (1) handwritten rules, (2) hand-coded features and Machine Learning, or (3) an End2End neural architecture.

Following the GREC evaluation campaign (Belz et al. 2009), in which a number of ML models of REG-in-Context were tested, a wider range of models were recently compared in terms of their performance, looking at human judgments and computational metrics (Same, Chen, and Van Deemter 2022). Models included (1) two rule-based ones, RREG-S (small) and RREG-L (small); (2) two models based on traditional Machine Learning (ML), called ML-S (small) and ML-L (large); and (3) three neural models, including two from Cunha et al. (2020) and one from Cao and Cheung (2019).

Comparison 1 and 2: Performance and direct support. Having observed that neural models had only been tested on Ferreira et al. (2018)’s version of WebNLG, Same, Chen, and Van Deemter (2022) decided to test all models on WSJ, the Wall Street Journal portion of the OntoNotes corpus (Gardent et al. 2017), arguing that WSJ would pose a better test for REG-in-Context algorithms because the texts in it are longer than those in WebNLG. With respect to WSJ, ML-L outperformed all other models; the simplest rule-based baseline RREG-S performed less well yet it performed at least as well as the neural models on both corpora.

Comparison 3: Indirect support. Indirect support varied widely across models, with the larger models receiving the most support from the linguistics literature. RREG-L, for instance, rests on notions such as local focus (Brennan 1995) and syntactic parallelism (Henschel, Cheng, and Poesio 2000); the large ML model ML-L makes use of the grammatical role of the RE.

The question of Indirect support for neural models is debatable (see below). Unless these models are combined with probing (for the case of REG, see Chen, Same, and van Deemter (2021)) or other add-ons, it is difficult to link these models with linguistic insights. On the other hand, neural models may be more inherently cognitively plausible than models based on rules or on classical Machine Learning, because they are inspired by our knowledge of the human brain. Rather than either blithely rejecting or accepting this argument, this is one of those “higher principles” (section 2) that we should take seriously, while also rigorously investigating their validity (c.f., (Ritter et al. 2017)).

7 For the challenge of linking neural models with domain insights, see Kambhampati (2021).
Comparison 4: Parsimony. Although parsimony can be difficult to assess, some relevant comparisons are straightforward in this case. As observed in Same, Chen, and Van Deemter (2022), the two rule-based models only have the current and previous sentence available to them; the two ML-based models look at the current and all previous sentences; the three neural models have the entire text available to them. The two “large” models, RREGL and ML-L, contain more rules/features than their smaller counterpart and are consequently less parsimonious. The three models above were architecturally similar seq-2-seq models with attention in the style of Bahdanau, Cho, and Bengio (2014), which did not display any obvious differences in terms of parsimony.

4. Challenges in assessing the explanatory value of a model

Our case studies illustrate how a model may be superior in one respect but inferior in others. And although our first case study suggested a trade-off between parsimony and performance, in which researchers could “buy” an improvement in performance by sacrificing parsimony, the second case study suggests that this is not always the case.

On the other hand, some challenges have come to the fore as well, which I will briefly discuss here. I leave challenges surrounding performance aside here, because they have been widely discussed (e.g., Celikyilmaz, Clark, and Gao (2020); Sai, Mohankumar, and Khapra (2022); Reiter (2018); Alva-Manchego, Scarton, and Specia (2021)) about metrics; van der Lee et al. (2019) about human evaluation.

Direct support. When judging the direct support for a model, younger models tend to be harder to judge, because a younger model cannot be expected to have been subjected to as much scrutiny as an older one, limiting its opportunities for both negative and positive support. Where very different results are reported on different corpora (such as the WebNLG and WSJ corpus in Same, Chen, and Van Deemter (2022)), further research into the causes of the divergence are in order.

A further wrinkle in assessing direct support is that models are moving targets: when a model is examined for the second or third time, it is often a modified version of that initial model. What is really being assessed in such cases is not one model but a class of models or, to put it differently, the ideas underlying these models (e.g., that recency and animacy are factors in deciding between the type of RE.

Indirect support. Our discussion of rationality put a spotlight on the two “key questions” that govern indirect support (section 2). For if the relevant principle I, as invoked in support of a model, is the idea that behaviour is rational, then some evidence may be available for I; but, as it stands, I is too vague to offer strong support for the details of the model, because it does not follow from I that the discriminatory power of the properties in an RE needs to be maximised. Similarly, existing experimental insights do not by themselves dictate all the specifics of the PRO model.

Parsimony. The idea of parsimony is already well established in NLP practices such as induction of “causal” models (Geiger et al. 2021), knowledge distillation (Sanh et al. 2019), and pruning, where the idea is to get rid of parameters or layers that do not add to a model’s performance (Tessier 2021). Current practices in NLP do not typically involve systematic comparisons between models in terms of their parsimony, and performing such comparisons rigourously is far from trivial. Theoretical equipment for doing so is available in principle, however. Deterministic models, for instance, may be compared with each other in terms of their Kolmogorov complexity (Solomonoff 1960; Kolmogorov 1965); non-deterministic models may be compared in terms of their
Minimal Description Length (Solomonoff 1964; Gruenwald 2007; Voita and Titov 2020). Comparisons across different types of models seem more problematic; a complicating factor is that whereas traditional models tend to address one NLP task, neural “foundation” models such as BERT are adaptable to a wide variety of tasks, which would tend to make a direct comparison across the two types of model biased against foundation models.

5. Policy Implications

Rather than shying away from them, I believe that our community should embrace the research challenges entailed by an increased emphasis on explanatory value, and the debates that this will bring, including debates about alternative dimensions of explanatory value.

Based on my reasoning in the previous sections, I think it would be wrong to limit evaluation of models to only one aspect of their quality. There are parallels here with the assessment of people, where the influential DORA declaration suggests that the academic community should reduce its reliance on quantitative metrics. Just as academics can have different talents, the success of a model has different dimensions. In both cases, we should learn to juggle multiple dimensions and say things like, “Based on this experiment, model A has better performance than the older model B. Being relatively new, A still has lower levels of (positive and negative) direct support. However, A is more parsimonious and appears to have better indirect support than B.”

Policy-wise, researchers and reviewers should be encouraged to think about the explanatory value of models. Analogous to the ethics and limitations paragraphs that are now solicited by some NLP conferences, our community could encourage authors of conference papers to comment on all dimensions of explanatory value. Similar moves could be made by institutions that offer funding for scientific research: analogous to letting proposers discuss societal and economic impact, they could be asked to discuss parsimony, and both kinds of support, as well. Alternatively, reviewers could be urged to check these dimensions, similar to when reviewers are expected to look out for analyses of statistical significance wherever these are appropriate to the work submitted.

6. Conclusion

It is widely accepted that performance alone does not make a good model, because constructing and training models may require a lot of effort; because of concerns over energy consumption; and because of concerns over linguistic, ethnic and other biases (Bender et al. 2021). Likewise, novelty and applicability of a model can be important.

In this Squib, I have argued that another set of dimensions is of crucial importance, particularly when NLP arises from a scientific interest in the world around us; these dimensions, variants of which have often been discussed in connection with other sciences, attempt to make explicit what it means for a model to explain data. Furthermore, I have argued that explanatory value means more than only performance, and that explanatory value does not equal explainability.

---

8 The San Francisco Declaration On Research Assessment can be found on https://sfdora.org/read/.
9 For ethics paragraphs see https://2021.aclweb.org/ethics/Ethics-FAQ/ For limitations paragraphs, https://aclweb.org/portal/content/empirical-methods-natural-language-processing-emnlp-2022.
The borderlines of our discussion are debatable. For example, it can be argued that similar arguments apply to NLP-as-Engineering as well. After all, unwieldy models are difficult to maintain and update; models that lack support may fail to generalize, and risk having to be completely redesigned whenever customers’ requirements change. Furthermore, trends in NLP research tend to reflect wider tendencies; accordingly, explanation is taking a backseat in other areas of Artificial Intelligence as well (Levesque 2014; Kambhampati 2021). My conjecture is that the same dimensions of explanatory value, and similar implications for research method and policy, apply there as well.
7. Citations

References

Agrawal, Harsh, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8948–8957.

Akmajan, Adrian and Frank Heny. 1975. Introduction to the Principles of Transformational Syntax. MIT Press, Cambridge, Mass.

Alva-Manchego, F., C. Scarton, and L. Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. Computational Linguistics, 47(4):861–889.

Ariel, Mira. 1990. Accessing Noun-Phrase Antecedents. Routledge.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Belke, Eva and Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. European Journal of Cognitive Psychology, 14(2):237–266.

Belz, Anja, Eric Kow, Jette Viethen, and Albert Gatt. 2009. Generating referring expressions in context: The grec task evaluation challenges. In Proceedings of ENLG 2009, pages 294–327, Association for Computational Linguistics.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21, page 610–623, Association for Computing Machinery, New York, NY, USA.

Berkovitz, Joseph. 2008. Action at a Distance in Quantum Mechanics. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy, Winter 2008 edition. Metaphysics Research Lab, Stanford University.

Brennan, Susan E. 1995. Centering attention in discourse. Language and Cognitive processes, 10(2):137–167.

Brody, Michael. 1995. Lexico-logical form. MIT Press, Cambridge, Mass.

Cao, Meng and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3163–3172, Association for Computational Linguistics, Hong Kong, China.

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. CoRR, abs/2006.14799.

Chen, Guanya, Fahime Same, and Kees van Deemter. 2021. What can neural referential form selectors learn? In Proceedings of the 14th International Conference on Natural Language Generation, pages 154–166, Association for Computational Linguistics, Aberdeen, Scotland, UK.

Craver, Carl and James Tabery. 2019. Mechanisms in Science. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy, Summer 2019 edition. Metaphysics Research Lab, Stanford University.

Cunha, Rossana, Thiago Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. pages 2261–2272.

Dale, Robert. 1989. Cooking up referring expressions. In Proceedings of the 27th annual meeting on Association for Computational Linguistics (ACL’89), pages 68–75, Association for Computational Linguistics, Vancouver, BC.

Dale, Robert and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19(2):233–263.

van Deemter, Kees. 2016. Computational Models of Referring: A study in cognitive science. MIT Press, Cambridge, MA.

Degen, Judith, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to “overinformative” referring expressions. Psychological Review, 127(4):591.

Elster, Jon. 1983. Sour Grapes: studies in the subversion of rationality. MIT Press, Cambridge, MA.
Ferreira, Thiago Castro, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the webnlg corpus. In Proceedings of the 11th International Conference on Natural Language Generation, page 171–176. Association for Computational Linguistics, Tilburg University.

Fitzpatrick, Simon. 2022. Simplicity in the philosophy of science. In Internet Encyclopedia of Philosophy, ISSN 2161-0002.

Frank, Michael C. and Noah Goodman. 2016. Pragmatic Language Interpretation as Probabilistic Inference. Trends in Cognitive Sciences, 20(11):818–829.

Frank, Michael C and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. Science (New York, N.Y.), 336(6084):998.

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planners. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, page 179–188, Association for Computational Linguistics, Vancouver.

Gatt, A, E. Krahmer, R.P.G. van Gompel, and K. van Deemter. 2013. Factors causing overspecification in definite descriptions. In Proceedings of the 35th Annual Meeting of the Cognitive Science Society.

Gatt, Albert and Anya Belz. 2010. Introducing Shared Tasks to NLG: the TUNA Shared Task Evaluation Challenges. In Emiel Krahmer and Mariet Theune, editors, Empirical Methods in Natural Language Generation. Springer.

Geiger, Atticus, Hanson Lu, Thomas F Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In Advances in Neural Information Processing Systems.

Gershman, S.J., E.J. Horvitz, and J.B. Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 49:273–278.

Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health, 3:745–750.

Gigerenzer, Gerd and Reinhard Selten. 2002. Bounded Rationality. MIT Press, Cambridge, MA.

Van Gompel, Roger PG, Kees van Deemter, Albert Gatt, Rick Snoeren, and Emiel J Krahmer. 2019. Conceptualization in reference production: Probabilistic modeling and experimental testing. Psychological review, 126(3):345.

Greene, B. 2000. The elegant universe: Superstrings, hidden dimensions, and the quest for the ultimate theory. American Journal of Physics, 68(2):199–200.

Gruenwald, Peter. 2007. The Minimum Description Length Principle. MIT Press.

Gundel, Jeanette K, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. Language, pages 274–307.

Harnad, Stevan. 1989. Minds, machines and searle. Journal of Theoretical and Experimental Artificial Intelligence, (1):5–25.

Hempel, Carl G. and Paul Oppenheim. 1965. Studies in the logic of explanation. Philosophy of Science, 15(2):135–175.

Henschel, Renate, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In Proceedings of the 18th conference on Computational linguistics-Volume I, pages 306–312, Association for Computational Linguistics.

Hepburn, Brian and Hanne Andersen. 2021. Scientific Method. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Hodosh, Micah, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47:853–899.

Hossenfelder, Sabine. 2018. Lost in Math; How Beauty Leads Physics Astray. Basic Books, New York.

Hutter, M., S. Legg, and P. M.B. Vitanyi. 2007. Algorithmic probability. Scholarpedia, 2(8):2572. Revision #151509.

Jaynes, E.T. 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK.

Kahneman, D. and A. Tversky. 1982. Prospect theory: An analysis of decision under risk. In Handbook of the fundamentals of financial decision making.

Kambhampati, Subbharao. 2021. Polanyi’s revenge and ai’s new romance with tacit knowledge. Communications of the ACM, 64(2):31–32.

Kolmogorov, A.N. 1965. Three approaches to the quantitative definition of information. Problems Inform. Transmission, 1(1):1–7.

Koolen, Ruud, Albert Gatt, Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions.
Krahmer, Emiel and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Association for Computational Linguistics, Tokyo, Japan.

Lei, Tao, Regina Barzilay, and Tommi Jaakola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Levesque, H.J. 2014. On our best behaviour. *Artificial Intelligence*, 212:27–35.

Lombrozo, Tania. 2006. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

Luo, Ruotian and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

McKeown, Kathleen. 2020. Rewriting the past: Assessing the field through the lens of language generation. (keynote). In *Proceedings of the 7th European Workshop on Natural Language Generation*.

Overton, James A. 2013. “explain” in scientific discourse. *Synthese*, (190):1383–1405.

Popper, Karl. 1934. *Logik der Forschung*, Translated as “The Logic of Scientific Discovery”. Hutchinson, London, 1959.

Reiter, Ehud. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of (ACL 2020)*.

Ribeiro, M.T., S. Singh, and C. Guestrin. 2016. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*.

Ritter, Samuel, David G. T. Barrett, Adam Santoro, and Matt M. Botvinick. 2017. Cognitive psychology for deep neural networks: A shape bias case study.

Rubio-Fernandez, Paula. 2021. Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. *Humanities and Social Sciences Communications*, 8(1):1–15.

Sai, Ananya B., Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Same, Fahime, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: A re-evaluation of neural referring expression generation systems. In *Proceedings of ACL 2022*, Association for Computational Linguistics.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

Simon, Herbert. 1991. Bounded rationality and organizational learning. *Organisational Science*, 2.

Solomonoff, Ray J. 1964. A formal theory of inductive inference: part i. *Information and Control*, 7(1):1–22.

Solomonoff, R.J. 1960. *A preliminary report on a general theory of inductive inference*. Technical Report ZTB-138, Zator Company, Cambridge, Mass.

Tessier, Hugo. 2021. Neural network pruning 101, https://towardsdatascience.com/neural-network-pruning-101-af816aaea61.

Voita, Elena and Ivan Titov. 2020. Information-theoretic probing with minimum description length.

Woodward, James and Lauren Ross. 2021. Scientific explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Springer, pages 264–293.

Yu, Licheng, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions.