Human Language Understanding & Reasoning

Christopher D. Manning

The last decade has yielded dramatic and quite surprising breakthroughs in natural language processing through the use of simple artificial neural network computations, replicated on a very large scale and trained over exceedingly large amounts of data. The resulting pretrained language models, such as BERT and GPT-3, have provided a powerful universal language understanding and generation base, which can easily be adapted to many understanding, writing, and reasoning tasks. These models show the first inklings of a more general form of artificial intelligence, which may lead to powerful foundation models in domains of sensory experience beyond just language.

When scientists consider artificial intelligence, they mostly think of modeling or recreating the capabilities of an individual human brain. But modern human intelligence is much more than the intelligence of an individual brain. Human language is powerful and has been transformative to our species because it gives groups of people a way to network human brains together. An individual human may not be much more intelligent than our close relatives of chimpanzees or bonobos. These apes have been shown to possess many of the hallmark skills of human intelligence, such as using tools and planning; moreover, they have better short-term memory than we do.[1] When humans invented language is still, and perhaps will forever be, quite uncertain, but within the long evolutionary history of life on Earth, human beings developed language incredibly recently. The common ancestor of prosimians, monkeys, and apes dates to perhaps sixty-five million years ago; humans separated from chimps perhaps six million years ago, while human language is generally assumed to be only a few hundred thousand years old.[2] Once humans developed language, the power of communication quickly led to the ascendancy of Homo sapiens over other creatures, even though we are not as strong as an elephant nor as fast as a cheetah. It was much more recently that humans developed writing (only a bit more than five thousand years ago), allowing knowledge to be communicated across distances of time and space. In just a few thousand years, this information-sharing mechanism took us from the bronze age to the smartphones of today. A high-fidelity

https://doi.org/10.1162/DAED_a_01905

code allowing both rational discussion among humans and the distribution of information has allowed the cultural evolution of complex societies and the knowledge underlying modern technologies. The power of language is fundamental to human societal intelligence, and language will retain an important role in a future world in which human abilities are augmented by artificial intelligence tools.

For these reasons, the field of natural language processing (NLP) emerged in tandem with the earliest developments in artificial intelligence. Indeed, initial work on the NLP problem of machine translation, including the famous Georgetown-IBM demonstration in 1954, slightly preceded the coining of the term "artificial intelligence" in 1956.[3] In this essay, I give a brief outline of the history of natural language processing. I then describe the dramatic recent developments in NLP coming from the use of large artificial neural network models trained on very large amounts of data. I trace the dramatic progress that has been made in building effective NLP systems using these techniques, and conclude with some thoughts on what these models achieve and where things will head next.

The history of natural language processing until now can be roughly divided into four eras. The first era runs from 1950 to 1969. NLP research began as research in machine translation. It was imagined that translation could quickly build on the great successes of computers in code breaking during World War II. On both sides of the Cold War, researchers sought to develop systems capable of translating the scientific output of other nations. Yet, at the beginning of this era, almost nothing was known about the structure of human language, artificial intelligence, or machine learning. The amount of computation and data available was, in retrospect, comically small. Although initial systems were promoted with great fanfare, the systems provided little more than word-level translation lookups and some simple, not very principled rule-based mechanisms to deal with the inflectional forms of words (morphology) and word order.

The second era, from 1970 to 1992, saw the development of a whole series of NLP demonstration systems that showed sophistication and depth in handling phenomena like syntax and reference in human languages. These systems included SHRDLU by Terry Winograd, LUNAR by Bill Woods, Roger Schank's systems such as SAM, Gary Hendrix's LIFER, and GUS by Danny Bobrow.[4] These were all hand-built, rule-based systems, but they started to model and use some of the complexity of human language understanding. Some systems were even deployed operationally for tasks like database querying.[5] Linguistics and knowledge-based artificial intelligence were rapidly developing, and in the second decade of this era, a new generation of hand-built systems emerged, which had a clear separation between declarative linguistic knowledge and its procedural processing, and which benefited from the development of a range of more modern linguistic theories.

However, the direction of work changed markedly in the third era, from roughly 1993 to 2012. In this period, digital text became abundantly available, and the compelling direction was to develop algorithms that could achieve some level of language understanding over large amounts of natural text and that used the existence of this text to help provide this ability. This led to a fundamental reorientation of the field around empirical machine learning models of NLP, an orientation that still dominates the field today. At the beginning of this period, the dominant modus operandi was to get hold of a reasonable quantity of online text – in those days, text collections were generally in the low tens of millions of words – and to extract some kind of model out of these data, largely by counting particular facts. For example, you might learn that the kinds of things people *capture* are fairly evenly balanced between locations with people (like a *city*, *town*, or *fort*) and metaphorical notions (like *imagination*, *attention*, or *essence*). But counts on words only go so far in providing language understanding devices, and early empirical attempts to learn language structure from text collections were fairly unsuccessful.[6] This led most of the field to concentrate on constructing annotated linguistic resources, such as labeling the sense of words, instances of person or company names in texts, or the grammatical structure of sentences in treebanks, followed by the use of supervised machine learning techniques to build models that could produce similar labels on new pieces of text at runtime.

The period from 2013 to present extended the empirical orientation of the third era, but the work has been enormously changed by the introduction of deep learning or artificial neural network methods. In this approach, words and sentences are represented by a position in a (several hundred- or thousand-dimensional) real-valued vector space, and similarities of meaning or syntax are represented by proximity in this space. From 2013 to 2018, deep learning provided a more powerful method for building performant models: it was easier to model longer distance contexts, and models generalized better to words or phrases with similar meanings because they could exploit proximity in a vector space, rather than depending on the identity of symbols (such as word form or part of speech). Nevertheless, the approach was unchanged in building supervised machine learning models to perform particular analysis tasks. Everything changed in 2018, when NLP was the first major success of very large scale *self-supervised* neural network learning. In this approach, systems can learn an enormous amount of knowledge of a language and the world simply from being exposed to an extremely large quantity of text (now normally in the billions of words). The method of self-supervision by which this is done is for the system to create from the text its own prediction challenges, such as successively identifying each next word in the text given the previous words or filling in a masked word or phrase in a text. By repeating such prediction tasks billions of times and learning from its mistakes, so the model does better next time given a similar textual context, general knowledge of a language and the world is

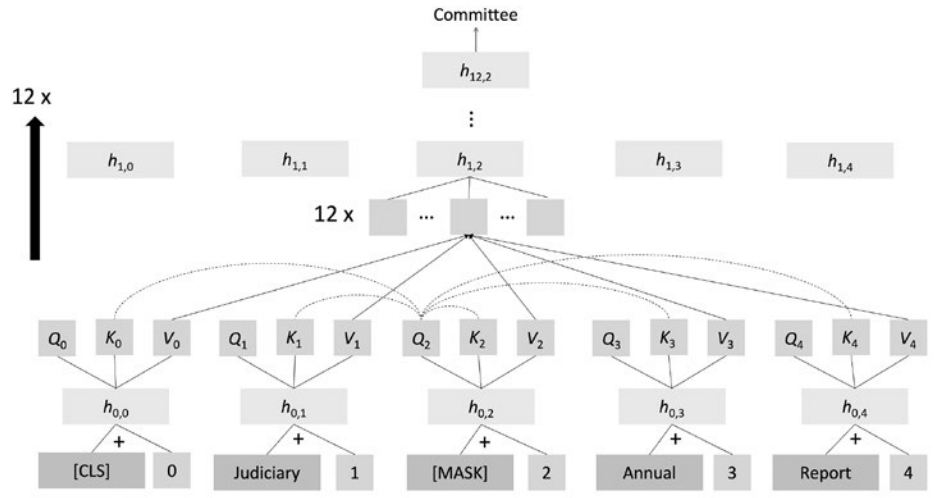
accumulated, and this knowledge can then be deployed for tasks of interest, such as question answering or text classification.

In hindsight, the development of large-scale self-supervised learning approaches may well be viewed as the fundamental change, and the third era might be extended until 2017. The impact of pretrained self-supervised approaches has been revolutionary: it is now possible to train models on huge amounts of unlabeled human language material in such a way as to produce one large pretrained model that can be very easily adapted, via fine-tuning or prompting, to give strong results on all sorts of natural language understanding and generation tasks. As a result, progress and interest in NLP have exploded. There is a sense of optimism that we are starting to see the emergence of knowledge-imbued systems that have a degree of general intelligence.

I cannot give here a full description of the now-dominant neural network models of human language, but I can offer an inkling. These models represent everything via vectors of real numbers and are able to learn good representations after exposure to many pieces of data by back-propagation of errors (which comes down to doing differential calculus) from some prediction task back to the representations of the words in a text. Since 2018, the dominant neural network model for NLP applications has been the transformer neural network.[7] With several ideas and parts, a transformer is a much more complex model than the simple neural networks for sequences of words that were explored in earlier decades. The dominant idea is one of *attention*, by which a representation at a position is computed as a weighted combination of representations from other positions. A common self-supervision objective in a transformer model is to mask out occasional words in a text. The model works out what word used to be there. It does this by calculating from each word position (including mask positions) vectors that represent a query, key, and value at that position. The query at a position is compared with the value at every position to calculate how much attention to pay to each position; based on this, a weighted average of the values at all positions is calculated. This operation is repeated many times at each level of the transformer neural net, and the resulting value is further manipulated through a fully connected neural net layer and through use of normalization layers and residual connections to produce a new vector for each word. This whole process is repeated many times, giving extra layers of depth to the transformer neural net. At the end, the representation above a mask position should capture the word that was there in the original text: for instance, *committee* as illustrated in Figure 1.

It is not at all obvious what can be achieved or learned by the many simple calculations of a transformer neural net. At first, this may sound like some kind of complex statistical association learner. However, given a very powerful, flexible, and high-parameter model like a transformer neural net and an enormous amount

Figure 1
Details of the Attention Calculations in One Part of a Transformer Neural Net Model



From this calculation, the transformer neural net is able to predict the word *committee* in the masked position.

of data to practice predictions on, these models discover and represent much of the structure of human languages. Indeed, work has shown that these models learn and represent the syntactic structure of a sentence and will learn to memorize many facts of the world, since each of these things helps the model to predict masked words successfully.[8] Moreover, while predicting a masked word initially seems a rather simple and low-level task – a kind of humorless Mad Libs – and not something sophisticated, like diagramming a sentence to show its grammatical structure, this task turns out to be very powerful because it is universal: every form of linguistic and world knowledge, from sentence structure, word connotations, and facts about the world, help one to do this task better. As a result, these models assemble a broad general knowledge of the language and world to which they are exposed. A single such large pretrained language model (LPLM) can be deployed for many particular NLP tasks with only a small amount of further instruction. The standard way of doing this from 2018 to 2020 was fine-tuning the model via a small amount of additional supervised learning, training it on the exact task of interest. But very recently, researchers have surprisingly found that the largest of these models, such as GPT-3 (Generative Pre-trained Transformer-3),

can perform novel tasks very well with just a *prompt*. Give them a human language description or several examples of what one wants them to do, and they can perform many tasks for which they were never otherwise trained.[9]

Traditional natural language processing models were elaborately composed from several usually independently developed components, frequently built into a pipeline, which first tried to capture the sentence structure and low-level entities of a text and then something of the higher-level meaning, which would be fed into some domain-specific execution component. In the last few years, companies have started to replace such traditional NLP solutions with LPLMs, usually fine-tuned to perform particular tasks. What can we expect these systems to do in the 2020s?

Early machine translation systems covered limited linguistic constructions in a limited domain.[10] Building large statistical models from parallel corpora of translated text made broad-coverage machine translation possible, something that most people first experienced using Google Translate after it launched in 2006. A decade later, in late 2016, Google's machine translation improved markedly when they switched to the use of neural machine translation.[11] But that system had a shorter lifespan: transformer-based neural translation was rolled out in 2020.[12] This new system improved not only via a different neural architecture but via use of a fundamentally different approach. Rather than building numerous pairwise systems from parallel text that translate between two languages, the new system gains from one huge neural net that was simultaneously trained on all languages that Google Translate covers, with input simply marked by a distinct token that indicates the language. While this system still makes mistakes and machine translation research continues, the quality of automatic translation today is remarkable. When I enter a couple of sentences from today's *Le Monde* culture section:

> *Il avait été surnommé, au milieu des années 1930, le « Fou chantant », alors qu'il faisait ses débuts d'artiste soliste après avoir créé, en 1933, un duo à succès avec le pianiste Johnny Hess. Pour son dynamisme sur scène, silhouette agile, ses yeux écarquillés et rieurs, ses cheveux en bataille, surtout pour le rythme qu'il donnait aux mots dans ses interprétations et l'écriture de ses textes.*[13]

the translation is excellent:

> *He was nicknamed the Singing Madman in the mid-1930s when he was making his debut as a solo artist after creating a successful duet with pianist Johnny Hess in 1933. For his dynamism on stage, his agile figure, his wide, laughing eyes, his messy hair, especially for the rhythm he gave to the words in his interpretations and the writing of his texts.*

In *question answering*, a system looks for relevant information across a collection of texts and then provides answers to specific questions (rather than just re-

turning pages that are suggested to hold relevant information, as in the early generations of Web search). Question answering has many straightforward commercial applications, including both presale and postsale customer support. Modern neural network question-answering systems have high accuracy in extracting an answer present in a text and are even fairly good at working out that no answer is present. For example, from this passage:

> Samsung saved its best features for the Galaxy Note 20 Ultra, including a more refined design than the Galaxy S20 Ultra – a phone I don't recommend. You'll find an exceptional 6.9-inch screen, sharp 5x optical zoom camera and a swifter stylus for annotating screenshots and taking notes. The Note 20 Ultra also makes small but significant enhancements over the Note 10 Plus, especially in the camera realm. Do these features justify the Note 20 Ultra's price? It begins at $1,300 for the 128GB version. The retail price is a steep ask, especially when you combine a climate of deep global recession and mounting unemployment.

One can get answers to questions like the following (using the UnifiedQA model):[14]

> How expensive is the Samsung Galaxy Note 20 Ultra?
> *$1,300 for the 128GB version*
>
> Does the Galaxy Note 20 Ultra have 20x optical zoom?
> *no*
>
> What is the optical zoom of the Galaxy Note 20 Ultra?
> *5x*
>
> How big is the screen of the Galaxy Note 20 Ultra?
> *6.9-inch*

For common traditional NLP tasks like marking person or organization names in a piece of text or classifying the sentiment of a text about a product (as positive or negative), the best current systems are again based on LPLMs, usually fine-tuned by providing a set of examples labeled in the desired way. While these tasks could be done quite well even before recent large language models, the greater breadth of knowledge of language and the world in these models has further improved performance on these tasks.

Finally, LPLMs have led to a revolution in the ability to generate fluent and connected text. In addition to many creative uses, such systems have prosaic uses ranging from writing formulaic news articles like earnings or sports reports and automating summarization. For example, such a system can help a radiologist by suggesting the impression (or summary) based on the radiologist's findings. For the findings below, we can see that the system-generated impression is quite similar to a radiologist-generated impression:[15]

> Findings: lines/tubes: right ij sheath with central venous catheter tip overlying the svc. on initial radiograph, endotracheal tube between the clavicular heads, and enteric tube with side port at the ge junction and tip below the diaphragm off the field-of-view; these are removed on subsequent film. mediastinal drains and left thoracostomy tube are unchanged. lungs: low lung volumes. retrocardiac airspace disease, slightly increased on most recent film. pleura: small left pleural effusion. no pneumothorax. heart and mediastinum: postsurgical widening of the cardiomediastinal silhouette. aortic arch calcification. bones: intact median sternotomy wires.
>
> Radiologist-generated impression: left basilar airspace disease and small left pleural effusion. lines and tubes positioned as above.
>
> System-generated impression: lines and tubes as described above. retrocardiac airspace disease, slightly increased on most recent film. small left pleural effusion.

These recent NLP systems perform very well on many tasks. Indeed, given a fixed task, they can often be trained to perform it as well as human beings, on average. Nevertheless, there are still reasons to be skeptical as to whether these systems really understand what they are doing, or whether they are just very elaborate rewriting systems, bereft of meaning.

The dominant approach to describing meaning, in not only linguistics and philosophy of language but also for programming languages, is a *denotational semantics* approach or a *theory of reference*: the meaning of a word, phrase, or sentence is the set of objects or situations in the world that it describes (or a mathematical abstraction thereof). This contrasts with the simple *distributional semantics* (or *use theory of meaning*) of modern empirical work in NLP, whereby the meaning of a word is simply a description of the contexts in which it appears.[16] Some have suggested that the latter is not a theory of semantics at all but just a regurgitation of distributional or syntactic facts.[17] I would disagree. Meaning is not all or nothing; in many circumstances, we partially appreciate the meaning of a linguistic form. I suggest that meaning arises from understanding the network of connections between a linguistic form and other things, whether they be objects in the world or other linguistic forms. If we possess a dense network of connections, then we have a good sense of the meaning of the linguistic form. For example, if I have held an Indian *shehnai*, then I have a reasonable idea of the meaning of the word, but I would have a richer meaning if I had also heard one being played. Going in the other direction, if I have never seen, felt, or heard a *shehnai*, but someone tells me that *it's like a traditional Indian oboe*, then the word has some meaning for me: it has connections to India, to wind instruments that use reeds, and to playing music. If someone added that *it has holes sort of like a recorder, but it has multiple reeds and a flared end more like an oboe*, then I have more network con-

nections to objects and attributes. Conversely, I might not have that information but just a couple of contexts in which the word has been used, such as: *From a week before, shehnai players sat in bamboo machans at the entrance to the house, playing their pipes. Bikash Babu disliked the shehnai's wail, but was determined to fulfil every conventional expectation the groom's family might have*.[18] Then, in some ways, I understand the meaning of the word *shehnai* rather less, but I still know that it is a pipe-like musical instrument, and my meaning is not a subset of the meaning of the person who has simply held a *shehnai*, for I know some additional cultural connections of the word that they lack.

Using this definition whereby understanding meaning consists of understanding networks of connections of linguistic forms, there can be no doubt that pretrained language models learn meanings. As well as word meanings, they learn much about the world. If they are trained on encyclopedic texts (as they usually are), they will learn that Abraham Lincoln was born in 1809 in Kentucky and that the lead singer of Destiny's Child was Beyoncé Knowles-Carter. Our machines can richly benefit from writing as a store of human knowledge, just like people. Nevertheless, the models' word meanings and knowledge of the world are often very incomplete and cry out for being augmented with other sensory data and knowledge. Large amounts of text data provided a very accessible way first to explore and build these models, but it will be useful to expand to other kinds of data.

The success of LPLMs on language-understanding tasks and the exciting prospects for extending large-scale self-supervised learning to other data modalities – such as vision, robotics, knowledge graphs, bioinformatics, and multimodal data – suggests exploring a more general direction. We have proposed the term *foundation models* for the general class of models with millions of parameters trained on copious data via self-supervision that can then easily be adapted to perform a wide range of downstream tasks.[19] LPLMs like BERT (Bidirectional Encoder Representations from Transformers) and GPT-3 are early examples of foundation models, but work is now underway more broadly.[20] One direction is to connect language models with more structured stores of knowledge represented as a knowledge graph neural network or as a large supply of text to be consulted at runtime.[21] However, the most exciting and promising direction is to build foundation models that also take in other sensory data from the world to enable integrated, multimodal learning. An example of this is the recent DALL·E model that, after self-supervised learning on a corpus of paired images and text, can express the meaning of a new piece of text by producing a corresponding picture.[22]

We are still very early in the era of foundation models, but let me sketch a possible future. Most information processing and analysis tasks, and perhaps even things like robotic control, will be handled by a specialization of one of a relatively small number of foundation models. These models will be expensive and time-consuming to train, but adapting them to different tasks will be quite easy;

indeed, one might be able to do it simply with natural language instructions. This resulting convergence on a small number of models carries several risks: the groups capable of building these models may have excessive power and influence, many end users might suffer from any biases present in these models, and it will be difficult to tell if models are safe to use in particular contexts because the models and their training data are so large. Nevertheless, the ability of these models to deploy knowledge gained from a huge amount of training data to many different runtime tasks will make these models powerful, and they will for the first time demonstrate the artificial intelligence goal of one machine learning model doing many particular tasks based on simply being instructed on the spot as to what it should do. While the eventual possibilities for these models are only dimly understood, they are likely to remain limited, lacking a human-level ability for careful logical or causal reasoning. But the broad effectiveness of foundation models means that they will be very widely deployed, and they will give people in the coming decade their first glimpses of a more general form of artificial intelligence.

ABOUT THE AUTHOR

Christopher D. Manning is the Thomas M. Siebel Professor in Machine Learning and Professor of Linguistics and of Computer Science at Stanford University; Director of the Stanford Artificial Intelligence Laboratory (SAIL); and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). He also served as President of the Association for Computational Linguistics. He is the author of *Introduction to Information Retrieval* (with Hinrich Schütze and Prabhakar Raghavan, 2008), *Foundations of Statistical Natural Language Processing* (with Hinrich Schütze, 1999), and *Complex Predicates and Information Spreading in LFG* (with Avery Andrews, 1999).

ENDNOTES

[1] Frans de Waal, *Are We Smart Enough to Know How Smart Animals Are?* (New York: W. W. Norton, 2017).

[2] Mark Pagel, "Q&A: What Is Human Language, When Did It Evolve and Why Should We Care?" *BMC Biology* 15 (1) (2017): 64.

[3] W. John Hutchins, "The Georgetown-IBM Experiment Demonstrated in January 1954," in *Machine Translation: From Real Users to Research*, ed. Robert E. Frederking and Kathryn B. Taylor (New York: Springer, 2004), 102–114.

[4] A survey of these systems and references to individual systems appears in Avron Barr, "Natural Language Understanding," *AI Magazine*, Fall 1980.

[5] Larry R. Harris, "Experience with Robot in 12 Commercial, Natural Language Data Base Query Applications" in *Proceedings of the 6th International Joint Conference on Artificial Intelligence, IJCAI-79* (Santa Clara, Calif.: International Joint Conferences on Artificial Intelligence Organization, 1979), 365–371.

[6] Glenn Carroll and Eugene Charniak, "Two Experiments on Learning Probabilistic Dependency Grammars from Corpora," in *Working Notes of the Workshop Statistically-Based NLP Techniques*, ed. Carl Weir, Stephen Abney, Ralph Grishman, and Ralph Weischedel (Menlo Park, Calif.: AAAI Press, 1992).

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems* 30 (2017).

[8] Christopher D. Manning, Kevin Clark, John Hewitt, et al., "Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision," *Proceedings of the National Academy of Sciences* 117 (48) (2020): 30046–30054.

[9] Tom Brown, Benjamin Mann, Nick Ryder, et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

[10] For example, Météo translated Canadian weather reports between French and English; see Monique Chevalier, Jules Dansereau, and Guy Poulin, *TAUM-MÉTÉO: Description du système* (Montreal: Traduction Automatique à l'Université de Montréal, 1978).

[11] Gideon Lewis-Kraus, "The Great A.I. Awakening," *The New York Times Magazine*, December 18, 2016.

[12] Isaac Caswell and Bowen Liang, "Recent Advances in Google Translate," Google AI Blog, June 8, 2020, https://ai.googleblog.com/2020/06/recent-advances-in-google-translate.html.

[13] Sylvain Siclier, "A Paris, le Hall de la chanson fête les inventions de Charles Trenet," *Le Monde*, June 16, 2021, https://www.lemonde.fr/culture/article/2021/06/16/a-paris-le-hall-de-la-chanson-fete-les-inventions-de-charles-trenet_6084391_3246.html.

[14] Daniel Khashabi, Sewon Min, Tushar Khot, et al., "UnifiedQA: Crossing Format Boundaries with a Single QA System," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 1896–1907.

[15] Yuhao Zhang, Derek Merck, Emily Bao Tsai, et al., "Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 5108–5120.

[16] For an introduction to this contrast, see Gemma Boleda and Aurélie Herbelot, "Formal Distributional Semantics: Introduction to the Special Issue," *Computational Linguistics* 42 (4) (2016): 619–635.

[17] Emily M. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2020), 5185–5198.

[18] From Anuradha Roy, *An Atlas of Impossible Longing* (New York: Free Press, 2011).

[19] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv (2021), https://arxiv.org/abs/2108.07258.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019), 4171–4186.

[21] Robert Logan, Nelson F. Liu, Matthew E. Peters, et al., "Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, Pa.: Association for Computational Linguistics, 2019), 5962–5971; and Kelvin Guu, Kenton Lee, Zora Tung, et al., "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of Machine Learning Research* 119 (2020).

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al., "Zero-Shot Text-To-Image Generation," arXiv (2021), https://arxiv.org/abs/2102.12092.