CS447 Literature Review: Large Language Models and Natural Language Understanding

Brian Reinbold
brianjr3@illinois.edu

December 9, 2022

Abstract

This literature review explores what large language models like BERT actually learn, and how close it is to natural language understanding. By contemplating the subtleties of what is linguistic meaning, we will see that large language models are far from true natural language understanding. Still, understanding these shortcomings together with a more precise, nuanced understanding of linguistic meaning can help provide direction for future research. Although this review will focus on BERT, many of the points apply to any large language model that is trained on form alone.

1 Introduction

Developments in neural network models have led to significant progress in natural language processing. The introduction of transformers followed by the innovation of BERT resulted in state-of-the-art performance in many NLP tasks like inference, natural language understanding, and question-answering. Given the impressive results of large language models, researchers have begun to probe what do large language models actually learn, and how close is this to actual natural language understanding?

Large language models, like BERT, seem to learn something about word knowledge. These models easily learn complex, non-linear relations between words and can exploit the statistical distributional hypothesis of words to construct word knowledge. Also, these models seem to understand syntax, however, the structures they learn do not necessarily translate to human annotated syntax trees. As words and syntax capture some aspects of meaning, large language models can learn some semantic knowledge, but they struggle with general pragmatic knowledge.

One way to address these shortcoming is simply to improve upon these models and train models with more parameters. This approach will definitely improve results, however, it comes with diminishing marginal returns, i.e., a model that is twice as large, although will learn more, will not learn twice as much. Another approach is to take a step back and reflect on what is natural language understanding, and what aspects do large language models exhibit and fail at it? There is a line of reasoning that natural language understanding cannot be learned from form alone, thus large language models can never truly reach natural language understanding. Still, exploring the nuances behind linguistic meaning can help researchers better evaluate model performance, design more informative NLP tasks, and motivate future research directions.

As we will see, the key to reaching natural language understanding will rely on incorporating additional sources beyond form. Human language understanding comes from interaction in the

world and interpersonal communication, and meaning derives from form's function to reach some goal in the real world.

The paper is organized as follows. The first section will provide a brief overview of some key ideas for understanding how large language models work. The first paper reviewed, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding* provides a brief overview of the large language model that sparked tremendous innovation in NLP (Devlin et al., 2019). Then *A Primer in BERTology: What We Know About How BERT Works* gives an overview of the key things BERT learns (Rogers et al., 2020). Next, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data* explores the subtleties of natural language understanding and meaning (Bender and Koller, 2020). This paper is more abstract and philosophical than the previous two, but understanding the nuances of linguistic meaning that the authors present will help researchers better understand what a large language model is capable of and likely cannot achieve. Finally, *Experience Grounds Language* reinforces the ideas from the previous paper. It provides a roadmap of where NLP has been, where it is at, and directions researchers need to pursue in order to make progress in natural language understanding (Bisk et al., 2020).

2 Technical Background

Although this review is not too focused on the technical details of large language models, this section provides a brief overview of some key ideas that will be useful for understanding how BERT works. Models in NLP often incorporate neural networks, which pass an input through layers of nodes and non-linear activation functions to produce an output. Also, many models exploit the fact that natural language consists of a sequence of words, which implies that previous words can help predict the following words in a sentence. A simple model utilizing this observation is a recurrent neural network (RNN) that modifies a neural network for sequential data (Sherstinsky, 2018). Another idea used in certain NLP tasks, such as machine translation, is the encoder-decoder architecture found in a seq2seq model (Sutskever et al., 2014). One RNN can be used to build a representation of the source language, in other words it encodes the source language, and then a second RNN can take this encoding to decode the representation into the target language.

Transformers incorporate and improve on the above ideas. They also use an encoder-decoder structure. One drawback of the RNN architecture is that it is not conducive to parallelization because it has to process tokens sequentially. Transformers, on the other hand, use masked self-attention. Self-attention has access to the entire input sequence and calculates a probability distribution over the encoder's representation. This allows the model to focus on different parts of the input. For example, in the following sentence, the subject is far from the verb: "The cats, clawing at the door, are hungry." Simply looking at the previous word would lead to incorrect subject-verb agreement. Attention can put more weight on the beginning of the sentence to help with predictions in later parts of the sentence. Masking randomly hides some tokens in the decoder, so that the model is forced to learn to predict these missing tokens. Masking allows the entire sequence to be passed to the decoder simultaneously instead of sequentially, allowing greater parallelization. BERT and other large language models build off the ideas of transformers to deliver state-of-the-art performance in various NLP tasks.

3 BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding

The **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, or **BERT**, was a significant innovation in NLP that has yielded impressive results. It has achieved state-of-the-art performance in eleven tasks ranging from natural language understanding to question answering (Devlin et al., 2019). The introduction of BERT is also the point in which academics and popular media begin to claim that large language models "understand" human language (Bender and Koller, 2020).

3.1 BERT: Architecture

There are two phases to training BERT: pre-training and fine tuning. The goal of pre-training is to learn a general representation that can be utilized in downstream tasks. A key advantage of BERT is its architectural flexibility - the same architecture is used in pre-training and fine tuning. The only difference in the fine tuning phase is that the input and output representation is swapped out for task specific representations. The weights from the pre-training phase are used to initialize the model for the fine-tuning phase, and then the model is trained end-to-end.

The input/output representation can handle a single sentence and pairs of sentences seamlessly. Sentences are simply separated by the special token $[SEP]$. Words are embedded using WordPiece (Wu et al., 2016). Then the input representation is the sum of the token embeddings, the segment embeddings (encodes which sentence it belongs to), and the position embeddings (where in the sequence the token is).

BERT incorporates the multi-layer bidirectional Transformer encoder found in Vaswani et al. (2017). Their base model has 12 layers, a hidden size of 768, 12 self-attention heads, 110 million parameters; and their large model has 24 layers, a hidden size of 1,024, 16 self-attention heads, and 340 million parameters.

BERT is trained on two unsupervised tasks in pre-training: 1) masked language modeling and 2) next sentence prediction. In masked language modeling, several tokens are randomly hidden using the special token $[MASKED]$, and the model tries to predict the masked words. Since the $[MASKED]$ token only appears in pre-training and not in fine tuning, this masking could prevent the model from generalizing, but this issue is mitigated by also replacing a $[MASKED]$ token with a random word token instead. The key innovation with masked language modeling is it allows the pre-trained model to learn a bidirectional representation by jointly conditioning on the left and right context. Before BERT, large language models were often unidirectional, which can be very limiting since having access to both sides of the context can be useful for many task such as question answering.

The second task, next sentence prediction, simply takes a sentence, and half the time, it is paired with the actual next sentence, and the other half, a random sentence. The goal is learn if the next sentence follows or not. Although simple, the authors demonstrate that it is useful for learning the relationship between sentences, which is essential in tasks like question-answering.

3.2 Performance

BERT achieved state-of-art performance in the GLUE, SQuAD v1.1, SQuAD v2.0, and SWAG benchmarks. GLUE includes nine datasets to test various NLP tasks. They include classification tasks in sentiment analysis, determining if a sentence is well formed, whether two questions are semantically related, identifying sentence entailment, and several others. This benchmark assess

whether a model can generalize over a diverse set of NLP task instead of specializing in one narrow task (Wang et al., 2018). SQuAD is the Stanford Question Answering Dataset. Given a question and a passage, the goal is to identify the answer to the question in the provided passage. SQuAD v2.0 extends v1.1 by allowing for the possibility that the answer to the question cannot be found in the given paragraph, which is a more realistic task (Rajpurkar et al., 2016). Finally, the Situations with Adversarial Generations (SWAG) dataset includes sentence-pair completion tasks to pick the most plausible sentence continuation to test commonsense inference (Zellers et al., 2018). All in all, BERT achieved state-of-the-art performance in eleven NLP tasks.

3.3 Discussion

Given the performance on diverse NLP tasks, surely BERT is learning something about natural language. For example, it would be difficult to answer a question correctly without actually understanding the question. The next paper will move beyond just looking at task performance and attempt to answer what BERT learns.

Also, even though this literature review focuses on BERT, many of the findings and conclusions will apply to other large language models. BERT's variants, GPT-3, XLNet, and other large language models possess common architectural elements, test using similar benchmarks, and more importantly they are trained on similar forms of data (Brown et al., 2020; Yang et al., 2019). Although these models may have varying performance, they likely learn similar things.

4 A Primer in BERTology: What We Know About How BERT Works

The authors survey over 150 studies to understand better what BERT actually learns (Rogers et al., 2020). They review BERT's syntactic, word, and semantic knowledge.

BERT does learn syntactic information, but it does not directly translate to annotated linguistic resources. BERT learns representations that are hierarchical instead of linear, which is similar to a syntax tree. BERT's embeddings also encode parts of speech (Liu et al., 2019). For certain tasks, BERT also takes subject-predicate agreement into account (Goldberg, 2019). However, despite this semblance of syntactic understanding, it is insensitive to malformed inputs such as shuffling words in a sentence. BERT also struggles with negation. These limitations raise the question whether BERT actually understands syntax (Ettinger, 2020). Still, it is possible that BERT's knowledge of syntax is simply incomplete, or it does not have to solely rely on it to make predictions (Glavaš and Vulić, 2021).

With regards to word knowledge, BERT can extract some world knowledge that is has access to in training, but it struggles with general pragmatic inference and role-based knowledge. If certain attributes and properties are not explicitly mentioned, it will struggle with general world knowledge. For example, BERT knows that people can walk into houses and houses are big, but BERT cannot infer that houses are bigger than people (Forbes et al., 2019).

Ettinger (2020) finds that BERT has some knowledge of semantic roles. Also, BERT encodes information about entity types, relations, and semantic roles (Tenney et al., 2019). However, BERT struggles with numbers and cannot generalize from the training data (Wallace et al., 2019).

4.1 Discussion

Overall, BERT seems to learn something about syntax and word meaning. Since syntax and word knowledge capture some aspects of meaning, BERT possesses some semantic knowledge, but it

does not have general, pragmatic world knowledge unless this specific knowledge is found in the training data. Now it is possible that large language models can be further improved upon, and a larger model would likely learn more. Still, there could be diminishing marginal returns to this approach. However, as we will see with the next two papers, large language models will likely be incapable of learning linguistic meaning from form alone.

5 Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

In this paper, the authors attempt to cool the hype of large language models. Although they have had impressive results, claims that these models "understand," "comprehend," or have learned "meaning" are overblown. Part of this confusion stems from loose use and definitions of linguistic meaning. The authors focus on the nuances and the philosophical perspective of meaning to evaluate large language models' understanding (Bender and Koller, 2020).

5.1 What is Meaning?

The authors distinguish *form* "to be any observable realization of a language" (e.g. text) while *meaning* "to be the relation between the form and communicative intent." Communicative intent can be to share information with someone, ask someone to do something, or just to socialize. The key to communicative intent is that it is something outside of language like something in the real world, an abstract idea in the world, or even in the speakers imagination. More formally, the authors denote meaning to be the relation $M \subseteq E \times I$ where E is a linguistic expression and I is communicative intent.

Next, the authors distinguish between communicative intent and conventional meaning. They define "the conventional meaning of an expression is what is constant across all of its possible contexts of use." Essentially, conventional meaning is akin to a dictionary definition. Therefore, any linguistic system defines pairs of expressions with conventional meanings. Formally, a linguistic system is a relation $C \subseteq E \times S$, where E are expressions and S are conventional meanings. Note that E is found in both M and C.

Based on these distinctions of meaning, communicative intent, conventional meaning, and their relations; the meaning M of a conversation between two people is best understood as facilitated by the linguistic system C. The speaker has a communicative intent, i, and chooses an expression e with a conventional meaning s, which tries to communicate i given the current social context. The listener hears the expression e, determines its conventional meaning s and their knowledge and assumptions of the world as well as the speaker's mind to deduce the speaker's communicative intent i. These relations formalize the fact that human conversation requires active participation between speaker and listener (Clark, 1996).

Also, deducing communicative intent requires knowledge of the world beyond the form of expressions, and this knowledge often is assumed so goes unsaid. Therefore, the authors argue that "a model of natural language that is trained purely on form will not learn meaning: if the training data is only form, there is not sufficient signal to learn the relation M between that form and the non-linguistic intent of human language users, nor C between form and the standing meaning the linguistic system assigns to each form."

5.2 Thought Experiments

The authors now present several thought experiments to make their abstract ideas about meaning more concrete. The first one is called the octopus test. Suppose two people are stranded on two different islands, but they can communicate with each other through a telegraph. Now imagine a hyper-intelligent octopus comes along and listens in on the inhabitants' conversation. After some time, the octopus feels lonely and cuts the underwater cable to take the place of one of the inhabitants in the conversation. Would the octopus have learned enough about language to take the place of one of the speakers and be able to fool the other speaker? In other words, would the person realize that they are no longer speaking with the other inhabitant?

The authors now suppose that the inhabitant invents a coconut catapult and shares the news (recall that this person thinks that they are talking to the other inhabitant, but they are really speaking with the hyper-intelligent octopus). Because the octopus has only had form, it is unlikely that it actually understands what the speaker is describing. The octopus cannot actually build a replica of this catapult or even manipulate its components underwater. Still, the octopus may be able to keep up the charade with generic responses like, "Neat idea!" However, the authors point out that if the speaker in this situation does not become suspicious, it is likely because they are assuming that the octopus understands their intent, which is a reasonable assumption when speaking with another person. But this then implies that the speaker is *assuming* understanding instead of the octopus *actually* understanding.

The octopus could really struggle to reply if asked, "which would fly further: a feather, a coconut, or a rock the size of a coconut?" Not to say that this task is easy for a human to answer, but at the least, a human would easily be able to reason about the different physical properties of these objects. The octopus has never seen these objects nor interacted with them. Also, these objects appear in very different, unrelated contexts, which would compound the difficulty in answering this question simply from learning by form alone.

Another useful thought experiment that summarizes their ideas is could one learn a language simply by listening to the radio? In this experiment, one only has access to linguistic form, but there is zero context. The learner may associate that certain sounds frequently occur together, but this does not suggest that there is any understanding of meaning of these sounds. For example, the learner could memorize utterances that describe cats, and so could likely answer, "What is a cat?" by simply repeating these utterances.

Now if this were a large language model tasked with distinguishing pictures of cats and dogs, would it be sensible to predict that it could perform this task beyond chance simply from being trained on form alone? Clearly not, since there was no additional visual context to associate form with the features in these images. Ultimately, meaning is more than just linguistic form, but it also incorporates things outside of form.

5.3 Discussion

Large language models like BERT demonstrate strong performance on certain meaning-related tasks, but that does not imply BERT possesses natural language understanding. Of course form, such as word associations and syntax, capture some semblance of meaning, but there are aspects of meaning that lie outside of form, like social context. Without access to social context and world knowledge, a model can not learn meaning on form alone. When children learn a language, research shows that learning is grounded in the real world, facilitated by interactions with others (Baldwin, 1995). By distinguishing between conventional meaning, communicative intent, expressions of

linguistic form, and how they relate with each other; researchers can develop a more nuanced understanding of natural language understanding that can better motivate how to test the performance and understanding of large language models. Furthermore, future research can draw inspiration on what is meaning and how humans learn a language to develop more sophisticated models that incorporate data beyond linguistic form to get closer to natural language understanding.

6 Experience Grounds Language

The authors of this final paper would likely agree with the points made by the authors in the previous article. The authors make the point that "meaning does not arise from the statistical distribution of words, but from their use by people to communicate" (Bisk et al., 2020). Although large language models have been effective at certain tasks after being trained on text alone, progress in natural language understanding will be stymied by failing to incorporate the context of the physical world "and the social interactions it facilitates."

The authors propose a framework of World Scopes to evaluate the progress of natural language understanding and verify if current research is headed in the right direction. These World Scopes go beyond simply understanding the meaning in text, but they also assess the contextual foundations of language, such as social interactions. These are their five World Scopes:

1. Corpus (past)
2. Internet (most of current NLP)
3. Perception (multimodal NLP)
4. Embodiment
5. Social

6.1 World Scope 1: Corpora and Representations

WS1 is emblematic of the datafication of natural language research. It includes annotated corpora, word vector representations, and embeddings. It has long been understood that context captures tremendous information about the meaning of words. For example, John Rupert Firth in his 1957 work, *A Synopsis of Linguistic Theory, 1930-1955*, is quoted saying, "You shall know a word by the company it keeps" (Firth, 1957).

Although not mentioned in this paper, Salton (1971) defined the term-document matrix that established the utility of vector representation for information retrieval. Vector representations indicate the presence or frequency of a word in a vocabulary set. Vectors numerically capture that co-occurrences and co-locations of words represent meaning, and two sentences are similar if their vector representations are similar. For example, although cats and dogs are quite different animals, they are both common house pets, so it is likely that these words will be used in similar sentences, for example, "I fed my cat/dog." Elman (1990) demonstrated that vector representations in an early neural network model could capture syntax and meaning. Furthermore, Elman showed that word meanings form hierarchical relationships by measuring the similarities of hidden unit activations.

Vector representations form sparse vectors, and since the number of dimensions of the vector is the size of the vocabulary, they suffer from the curse of dimensionality. Dense vector embeddings improve on vector representations by "learning a distributed representation for words" that significantly reduces the number of dimensions (Bengio et al., 2003). Embeddings are quicker to train and can also capture context better than sparse vectors. Dense vector embeddings exploded in

popularity with the introduction of the word2vec algorithm and the proliferation of deep learning in NLP (Mikolov et al., 2013).

Beyond vector representations of words, annotated corpora captured semantic and syntactic representations. Finding formal linquisitic structures could be used for recovering syntax trees. The Penn Treebank is the canonical example of this direction (Marcus et al., 1993).

Overall, representing words with other words is sensible since this is exactly how dictionaries and thesauruses relate words with their definitions and synonyms. However, it is not intuitive how to interpret the meaning of representations in deep learning language models. Still, large language models have made it possible to learn more complex text-based tasks using larger, unstructured text data.

6.2 World Scope 2: The Written World

This World Scope is characterized by the use of unstructured, unlabeled, multi-domain, and multilingual data that was made possible with the advent of the internet and large web crawlers. Building off of the previous World Scope, large language models have had impressive results, capturing rich syntactic structure and semantics that were not previously thought possible. Importantly, these advances have occurred in an unsupervised manner; large language models learn these representations themselves without the input of humans. No longer was progress constrained to a single source or annotation. Impressively, these general representations that these models learn can be transferred to diverse contexts (Brown et al., 2020). However, these gains come at the cost of training large models. Also, these model exhibit diminishing marginal returns.

Still, they are modeling the written world like what came before. No matter the scale, symbolic search of linguistic co-occurrences is devoid of meaning. Although their representations capture some aspects of meaning, it is still an open question as to how much of contextual understanding they possess. For example, these text only models still struggle with a sentence like this: "I parked my car in the compact parking space because it looked (big/small) enough." Ultimately, can one learn a language simply from listening to the radio?

6.3 World Scope 3: The World of Sights and Sounds

In the next World Scope, the authors argue that language needs perception, i.e., auditory, tactile, and visual input. Human perception of the world directly shapes language, and language is used to communicate about the world. For example, the metaphor, "as nimble as a cat," captures the observation that a falling cat lands quietly on its feet (Lakoff and Johnson, 2008). Furthermore, there is evidence that children require sensory perception to learn a language, suggesting that language understanding is more than just speech or words in a text (Vigliocco et al., 2014).

Research in computer vision has made great progress into complex visual and physical phenomena. For example, Mottaghi et al. (2016) construct a model that can reason about how forces act on objects. The authors also note that, "advances in computer vision have enabled building semantic representations rich enough to interact with natural language." For example, Mitchell et al. (2012) developed a model that can generate descriptive annotations of images. Further research can look to integrate advances in computer vision and NLP to build more intelligent systems that can incorporate perception into their understanding.

A model in WS3 would then be able to determine the correct word in the sentence from WS2, "I parked my car in the compact parking space because it looked ***small*** enough." However, the authors provide another example that this World Scope could not answer, "Would a ceramic or paper plate

make a better frisbee?" Answering this question requires interaction in the actual world, so a model in WS3 would fail the octupus test.

6.4 World Scope 4: Embodiment and Action

WS4 takes perception to the next level by including action-oriented interactions. Research in development psychology suggests that "intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity" (Smith and Gasser, 2005). Children acquire a language by manipulating the world around them and connecting their perceptions with their interactions. Language allows one to connect words with these actions. For example, humans learn the physical properties of objects through interaction and can then easily reason over different representations elicited by objects. The challenge in NLP is that much of this knowledge is implicitly understood and so goes unsaid. Therefore, large language models will not be able to learn these physical representations and interactions from text alone.

The authors provide an example that summarizes well how the different World Scopes would handle this question: "Is an orange more like a baseball or a banana?" WS1 may understand that they are all common nouns and possibly that oranges and bananas are fruits but likely not much else. WS2 may capture that oranges and baseballs both roll, but no additional physical properties like relative size and surface texture. WS3 may learn the relative deformability of these objects, but probably will not understand that baseballs are interacted with with much greater force. WS4 can appreciate the subtleties of the question and seamlessly switch between different representations of the objects. The orange and baseball can be manipulated similarly since they have similar shapes and sizes while the orange and banana both contain peels, deform, and are edible (Bisk et al., 2020).

Understanding interactions in the physical world moves beyond just computer vision since it requires the agent to actually be in the world. Innovations in robotics coupled with NLP could lead the way forward. ALFRED (Action Learning From Realistic Environments and Directives), presents a benchmark for learning how to map from natural language instructions and computer vision to sequences of action-oriented household tasks (Shridhar et al., 2019).

6.5 World Scope 5: The Social World

Finally, WS5 is about interpersonal communication, which is the primary use case of natural language. The previous World Scopes provide more sources of data, but the goal with this World Scope is to allow language to also be a *cause* - "to generate language that does something in the world." Instead of just being capable of regurgitating what has been previously seen, a model in this World Scope could create new, original thoughts, and it would be able to respond appropriately given the social context.

Conversation between humans is always grounded in social context. There are countless things that can be relevant to social context that is left unsaid, and which could not easily be captured by annotating datasets. As discussed in *Towards Natural Language Understanding*, conversation requires active participation between speaker and hearer, and it requires them to seamlessly interpret expressions with their conventional meanings to deduce speaker intent and thus derive meaning. The previous World Scopes can definitely help with this by allowing a model to learn varied representations of objects and ideas. However, WS5 takes it further by requiring the ability to reason about the speaker's state of mind. "The ability to consider the feelings and knowledge of others is now commonly referred to as the the 'Theory of Mind'" (Nematzadeh et al., 2018). The simple training-testing paradigm will not be sufficient to reach this World Scope, but one possibility may

be to have models and humans interact, and thus have the model continuously learn through these interactions.

6.6 Discussion

Further NLP progress seems daunting after seeing that there are five World Scopes, and large language models are only at WS2. However, WS3 and WS4 can probably be reached by improving current deep learning techniques. Deep learning models have very flexible architectures that can be applied to diverse data, and different representations can easily be combined. Computer vision has already made progress in areas like semantic representation. It seems likely it would not be a huge leap to combine insights from NLP, computer vision, and perhaps incorporate auditory data to progress towards WS3 and WS4. At the least, the foundation may already be there. Still, WS4 would require further advancements in robotics to incorporate additional sensorimotor data.

Finally, WS5 seems almost like science fiction at the moment since it may require sentient AI to be able to truly understand human emotions, social context, and generate new insights. Perhaps training utilizing reinforcement learning with constant human interaction providing immediate feedback could make tremendous progress. Still, AI may not need to completely reach these World Scopes to be perceived by humans as in these World Scopes due to the ELIZA effect (Weizenbaum, 1966). If AI shows a semblance of additional contextual knowledge, maybe humans will think that the AI is sentient anyways.

7 Discussion and Conclusion

This literature review explored what do large language models like BERT actually learn and how close is this to natural language understanding. Large language models, like BERT, seem to learn something about word, syntactic, and semantic knowledge. These features of language capture some aspects of linguistic meaning but are not the whole picture, and large language models can fail to generalize on pragmatic inferences that are not explicity trained on.

The key to natural language understanding is that "function is the source of meaning" (Wittgenstein, 1953). As discussed in *Climbing Towards Natural Language Understanding*, human language understanding comes from interaction in the world and interpersonal communication. Meaning derives from form's function to reach some goal in the real world. Importantly, aspects of meaning like communicative intent lie outside of form and thus cannot be learned from form alone. The direction forward then is for models to incorporate additional sources outside of text from the internet, such as images, videos, auditory signals, sensorimotor information, and interactions with humans.

Now one counterpoint to some of the ideas about what is natural language understanding, is that it relies on observing how humans learn a language. How humans learn can definitely provide insights, but computers are not humans. Maybe natural language understanding is a very human phenomena so cannot be achieved by computers, or maybe there will be other ways to arrive at it. Also, AI may only need to approximate natural language understanding to be perceived as possessing natural language understanding given the ELIZA effect (Weizenbaum, 1966). Perhaps some researchers are already under the spell of the ELIZA effect and believe their models "understand." Ultimately, exploring the nuances of linguistic meaning can help researchers better evaluate how close large language models are to natural language understanding and motivate future research.

References

Dare A. Baldwin. 1995. Understanding the link between joint attention and language.

Emily M. Bender and Alexander Koller. 2020. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

H.H. Clark. 1996. *Using Language*. ACLS Humanities E-Book. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

J.R. Firth. 1957. *A Synopsis of Linguistic Theory, 1930-1955*.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *CoRR*, abs/1908.02899.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *ArXiv*, abs/1901.05287.

G. Lakoff and M. Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander C. Berg, Tamara L. Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Kumar Gupta, and Ali Farhadi. 2016. "what happens if..." learning to predict the effect of forces in images. *ArXiv*, abs/1603.05600.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

G. Salton. 1971. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., USA.

Alex Sherstinsky. 2018. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2019. Alfred: A benchmark for interpreting grounded instructions for everyday tasks.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2):13–29.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *CoRR*, abs/1905.06316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*, 4th edition.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.