CREATE QUESTIONS

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

**Emily M. Bender**
University of Washington
Department of Linguistics
ebender@uw.edu

**Alexander Koller**
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

## Abstract

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has *a priori* no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

## 1 Introduction

The current state of affairs in NLP is that the large neural language models (LMs), such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019), are making great progress on a wide range of tasks, including those that are ostensibly meaning-sensitive. This has led to claims, in both academic and popular publications, that such models "understand" or "comprehend" natural language or learn its "meaning". From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning.

We argue that *the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning*. We take the term *language model* to refer to any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not. We take (linguistic) *meaning* to be the relation between a linguistic form and communicative intent.

Our aim is to advocate for an alignment of claims and methodology: Human-analogous natural language understanding (NLU) is a grand challenge of artificial intelligence, which involves mastery of the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that genuine progress in our field — climbing the right hill, not just the hill on whose slope we currently sit — depends on maintaining clarity around big picture notions such as *meaning* and *understanding* in task design and reporting of experimental results.

After briefly reviewing the ways in which large LMs are spoken about and summarizing the recent flowering of "BERTology" papers (§2), we offer a working definition for "meaning" (§3) and a series of thought experiments illustrating the impossibility of learning meaning when it is not in the training signal (§4,5). We then consider the human language acquisition literature for insight into what information humans use to bootstrap language learning (§6) and the distributional semantics literature to discuss what is required to ground distributional models (§7). §8 presents reflections on how we look at progress and direct research effort in our field, and in §9, we address possible counterarguments to our main thesis.

## 2 Large LMs: Hype and analysis

Publications talking about the application of large LMs to meaning-sensitive tasks tend to describe the models with terminology that, if interpreted at face value, is misleading. Here is a selection from academically-oriented pieces (emphasis added):

(1) In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task. (Devlin et al., 2019)

(2) Using BERT, a pre-training language model, has been successful for single-turn machine **comprehension** ... (Ohsugi et al., 2019)

(3) The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demon-

could it also be the nature of the beast which causes overclaims? Quote of Rise+fall of AI hype

Eg academics need to hype their research in order to get grants & draw public interest?

When is it simply hype vs. AI pseudoscience?

and not overexaggerated selling snake oil

which the public is wont to do.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

strates their potential as unsupervised open-domain QA systems. (Petroni et al., 2019)

If the highlighted terms are meant to describe human-analogous understanding, comprehension, or recall of factual knowledge, then these are gross overclaims. If, instead, they are intended as technical terms, they should be explicitly defined.

– One important consequence of imprudent use of terminology in our academic discourse is that it feeds AI hype in the popular press. As NLP gains public exposure and is more widely used in applied contexts, it is increasingly important that the actual capabilities of our systems be accurately represented. In some cases, NLP experts speaking with the media are being appropriately careful, as in these two quotes in the *New York Times*:[1]

(4) These systems are still a really long way from truly understanding running prose. (Gary Marcus)

(5) Though BERT passed the lab's common-sense test, machines are still a long way from an artificial version of a human's common sense. (Oren Etzioni)

However, there are plenty of instances where the popular press gets it wrong, such as (6) from the B2C website,[2] apparently based on the Google Blog post about BERT and search, which includes numerous statements like (7).[3]

(6) BERT is a system by which Google's algorithm uses pattern recognition to better **understand** how human beings communicate so that it can return more relevant results for users.

(7) Here are some of the examples that showed up our evaluation process that demonstrate BERTs ability to **understand** the intent behind your search.

In sum, it is not clear from our academic literature whether all authors are clear on the distinction between form and meaning, but it is clear that the way we speak about what neural LMs are doing is misleading to the public.

Part of the reason for this tendency to use imprecise language may well be that we do not yet fully understand what exactly it is about language that the large LMs come to implicitly represent. Their success, however, has sparked a subfield ('BERTology') that aims to answer this question. The methodology of probing tasks (e.g. Adi et al., 2017; Ettinger et al., 2018) has been used to show that large LMs learn at least some information about phenomena such as English subject-verb agreement (Goldberg, 2019; Jawahar et al., 2019), constituent types, dependency labels, NER, and (core) semantic role types (again, all in English) (Tenney et al., 2019).[4] Hewitt and Manning (2019) find information analogous to unlabeled dependency structures in the word vectors provided by ELMo and BERT (trained on English). And of course it is well established that vector-space representations of words pick up word classes, both syntactic (POS, e.g. Lin et al., 2015) and semantic (lexical similarity, e.g. Rubenstein and Goodenough, 1965; Mikolov et al., 2013).

Others have looked more closely at the success of the large LMs on apparently meaning sensitive tasks and found that in fact, far from doing the "reasoning" ostensibly required to complete the tasks, they were instead simply more effective at leveraging artifacts in the data than previous approaches. Niven and Kao (2019) find that BERT's unreasonably good performance on the English Argument Reasoning Comprehension Task (Habernal et al., 2018) falls back to chance if the dataset is modified by adding adversarial examples that just negate one piece of the original, thus mirroring the distribution of lexical cues for each label. Similarly, McCoy et al. (2019) find that BERT's performance on the English Multi-genre Natural Language Inference dataset (Williams et al., 2018) is predicated on its ability to leverage syntactic heuristics involving overlap (of full constituents, subsequences, or simply bags of words). In a dataset carefully designed to frustrate such heuristics, BERT's performance falls to significantly below chance.

In this brief overview of BERTology papers we have highlighted both the extent to which there is evidence that large LMs can learn aspects of linguistic formal structure (e.g. agreement, dependency structure), and how their apparent ability to "reason" is sometimes a mirage built on leveraging artifacts in the training data (i.e. form, not meaning). Our contribution is an argument on theoretical grounds that a system exposed only to form in its training cannot in principle learn meaning.

## 3 What is meaning?

We start by defining two key terms: We take *form* to be any observable realization of language: marks

[1] https://www.nytimes.com/2018/11/18/technology/artificial-intelligence-language.html, accessed 2019/12/04

[2] https://www.business2community.com/seo/what-to-do-about-bert-googles-recent-local-algorithm-update-02259261, accessed 2019/12/04

[3] https://www.blog.google/products/search/search-language-understanding-bert/, accessed 2019/12/04

[4] But see Warstadt et al.'s (2019) cautionary note about how the methodology used for probing can influence the results.

on a page, pixels or bytes in a digital representation of text, or movements of the articulators.[5] We take *meaning* to be the relation between the form and something external to language, in a sense that we will make precise below.

### 3.1 Meaning and communicative intent

When humans use language, we do so for a purpose: We do not talk for the joy of moving our articulators, but in order to achieve some *communicative intent*. There are many types of communicative intents: they may be to convey some information to the other person; or to ask them to do something; or simply to socialize. We take *meaning* to be the relation $M \subseteq E \times I$ which contains pairs $(e, i)$ of natural language expressions $e$ and the communicative intents $i$ they can be used to evoke. Given this definition of meaning, we can now use *understand* to refer to the process of retrieving $i$ given $e$.

Communicative intents are about something that is *outside of language*. When we say *Open the window!* or *When was Malala Yousafzai born?*, the communicative intent is grounded in the real world the speaker and listener inhabit together. Communicative intents can also be about abstract worlds, e.g. bank accounts, computer file systems, or a purely hypothetical world in the speaker's mind.

Linguists distinguish communicative intent from *conventional* (or *standing*) meaning (Quine, 1960; Grice, 1968). The conventional meaning of an expression (word, phrase, sentence) is what is constant across all of its possible contexts of use. Conventional meaning is an abstract object that represents the communicative potential of a form, given the linguistic system it is drawn from. Each linguistic system (say, English) provides a relation $C \subseteq E \times S$, which contains pairs $(e, s)$ of expressions $e$ and their conventional meanings $s$.[6] The field of linguistic semantics provides many competing theories of what conventional meanings $s$ look like. For our purposes, we don't need to select among these theories; all we assume is that conventional meanings must have interpretations, such as a means of testing them for truth against a model of the world. Thus, like the meaning relation $M$, $C$ connects language to objects outside of language.

[5] In spoken languages, the primary articulators are the components of the vocal tract. In signed languages, they are principally the hands and face.

[6] We abstract away here from the facts that linguistic systems $C$ change over time and are only incompletely shared among different speakers. They are stable enough to function as rich signals to communicative intent.

Returning to the meaning relation $M$ from above, it is best understood as mediated by the relation $C$ of a linguistic system shared between two interlocutors. The speaker has a certain communicative intent $i$, and chooses an expression $e$ with a standing meaning $s$ which is fit to express $i$ in the current communicative situation. Upon hearing $e$, the listener then reconstructs $s$ and uses their own knowledge of the communicative situation and their hypotheses about the speaker's state of mind and intention in an attempt to deduce $i$.

This active participation of the listener is crucial to human communication (Reddy, 1979; Clark, 1996). For example, to make sense of (8) and (9) (from Clark, 1996, p.144), the listener has to calculate that *Napoleon* refers to a specific pose (hand inside coat flap) or that *China trip* refers to a person who has recently traveled to China.

(8) The photographer asked me to do a Napoleon for the camera.

(9) Never ask two China trips to the same party.

We humans are also very willing, as we will see in §4 below, to attribute communicative intent to a linguistic signal of a language we speak, even if the originator of the signal is not an entity that could have communicative intent.

To summarize, as we strive to understand how NLU tasks and system performance on those tasks relates to the bigger picture goals of building human-analogous natural language understanding systems, it is useful to distinguish cleanly between form, conventional meaning, and communicative intent. Furthermore, we should be careful not to confuse communicative intent with ground truth about the world, as speakers can of course be mistaken, be intentionally dissembling, etc.

We 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.

### 3.2 Meaning and intelligence

Meaning and understanding have long been seen as key to intelligence. Turing (1950) argued that a machine can be said to "think" if a human judge cannot distinguish it from a human interlocutor after having an arbitrary written conversation with

each. However, humans are quick to attribute meaning and even intelligence to artificial agents, even when they know them to be artificial, as evidenced by the way people formed attachments to ELIZA (Weizenbaum, 1966; Block, 1981).

This means we must be extra careful in devising evaluations for machine understanding, as Searle (1980) elaborates with his Chinese Room experiment: he develops the metaphor of a "system" in which a person who does not speak Chinese answers Chinese questions by consulting a library of Chinese books according to predefined rules. From the outside, the system seems like it "understands" Chinese, although in reality no actual understanding happens anywhere inside the system.

Searle's thought experiment begins from the premise that it is possible to manipulate forms well enough to be indistinguishable from a system that understands the meaning of the forms, reasons about it, and responds appropriately. We observe that much recent work in NLP claims to be building systems where not only the runtime system but in fact also the process for building it only has access to form. But language is used for communication about the speakers' actual (physical, social, and mental) world, and so the reasoning behind producing meaningful responses must connect the meanings of perceived inputs to information about that world. This in turn means that for a human or a machine to learn a language, they must solve what Harnad (1990) calls the *symbol grounding problem*. Harnad encapsulates this by pointing to the impossibility for a non-speaker of Chinese to learn the meanings of Chinese words from Chinese dictionary definitions alone.

Our purpose here is to look more deeply into why meaning can't be learned from linguistic form alone, even in the context of modern hardware and techniques for scaling connectionist models to the point where they can take in vast amounts of data. We argue that, independently of whether passing the Turing test would mean a system is intelligent, a system that is trained only on form would fail a sufficiently sensitive test, because it lacks the ability to connect its utterances to the world.

## 4 The octopus test

In order to illustrate the challenges in attempting to learn meaning from form alone, we propose a concrete scenario. Say that A and B, both fluent speakers of English, are independently stranded on two uninhabited islands. They soon discover that previous visitors to these islands have left behind telegraphs and that they can communicate with each other via an underwater cable. A and B start happily typing messages to each other.

Meanwhile, O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations. O knows nothing about English initially, but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A's utterances. O also observes that certain words tend to occur in similar contexts, and perhaps learns to generalize across lexical patterns by hypothesizing that they can be used somewhat interchangeably. Nonetheless, O has never observed these objects, and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives.

At some point, O starts feeling lonely. He cuts the underwater cable and inserts himself into the conversation, by pretending to be B and replying to A's messages. Can O successfully pose as B without making A suspicious? This constitutes a weak form of the Turing test (weak because A has no reason to suspect she is talking to a non-human); the interesting question is whether O fails it because he has not learned the meaning relation, having seen only the form of A and B's utterances.

The extent to which O can fool A depends on the task — that is, on what A is trying to talk about. A and B have spent a lot of time exchanging trivial notes about their daily lives to make the long island evenings more enjoyable. It seems possible that O would be able to produce new sentences of the kind B used to produce; essentially acting as a chatbot. This is because the utterances in such conversations have a primarily social function, and do not need to be grounded in the particulars of the interlocutors' actual physical situation nor anything else specific about the real world. It is sufficient to produce text that is internally coherent.

Now say that A has invented a new device, say a coconut catapult. She excitedly sends detailed instructions on building a coconut catapult to B, and asks about B's experiences and suggestions for improvements. Even if O had a way of constructing the catapult underwater, he does not know what words such as *rope* and *coconut* refer to, and thus can't physically reproduce the experiment. He can

only resort to earlier observations about how B responded to similarly worded utterances. Perhaps O can recognize utterances about *mangos* and *nails* as "similarly worded" because those words appeared in similar contexts as *coconut* and *rope*. So O decides to simply say "Cool idea, great job!", because B said that a lot when A talked about ropes and nails. It is absolutely conceivable that A accepts this reply as meaningful — but only because A does all the work in attributing meaning to O's response. It is not because O understood the meaning of A's instructions or even his own reply.

Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.[7]

Having only form available as training data, O did not learn meaning. The language exchanged by A and B is a projection of their communicative intents through the meaning relation into linguistic forms. Without access to a means of hypothesizing and testing the underlying communicative intents, reconstructing them from the forms alone is hopeless, and O's language use will eventually diverge from the language use of an agent who can ground their language in coherent communicative intents.

The thought experiment also illustrates our point from §3 about listeners' active role in communication. When O sent signals to A pretending to be B, he exploited statistical regularities in the form, i.e. the distribution of linguistic forms he observed. Whatever O learned is a reflection of A and B's communicative intents and the meaning relation. But reproducing this distribution is not sufficient for meaningful communication. O only fooled A into believing he was B because A was such an active listener: Because agents who produce English sentences usually have communicative intents, she assumes that O does too, and thus she builds the conventional meaning English associates with O's utterances. Because she assumes that O is B, she uses that conventional meaning together with her other guesses about B's state of mind and goals to attribute communicative intent. It is not that O's utterances make sense, but rather, that A can make sense of them.

[7] To see what a large LM might reply in this situation, we prompted the GPT-2 demo with *"Help! I'm being chased by a bear! All I have is these sticks. What should I do?"*, and GPT-2 to supplied *"You're not going to get away with this!"* (https://gpt2.apps.allenai.org/, accessed 2019/12/4). Following Radford et al.'s (2019) approach of giving explicit cues to encode the task, we also constructed a more elaborate prompt. The results, given in Appendix A, are highly entertaining but no more helpful to the hapless A.

## 5 More constrained thought experiments

The story of the octopus considers the problem of learning not only the full communicative system, including the relations $M$ and $C$, but also the reasoning required to come up with answers that are both coherent and also helpful in the real world. Here, we provide two more constrained thought experiments, to focus more narrowly on the problem of learning the meaning relation, for both natural languages and programming languages.

Because programming languages are designed to be unambiguous and relatively insensitive to execution context, the distinction between standing and speaker meaning is less important than for natural languages. A Java program $e$, when compiled and executed on the Java Virtual Machine, can be interpreted as a function $i$ which maps program inputs to program outputs. We take the meaning relation $J \subseteq E \times I$ of Java to contain all such pairs $(e, i)$.

**Java** Imagine that we were to train an LM on all of the well-formed Java code published on Github. The input is only the code. It is not paired with bytecode, nor a compiler, nor sample inputs and outputs for any specific program. We can use any type of LM we like and train it for as long as we like. We then ask the model to execute a sample program, and expect correct program output.

**English** As as second example, imagine training an LM (again, of any type) on English text, again with no associated independent indications of speaker intent. The system is also given access to a very large collection of unlabeled photos, but without any connection between the text and the photos. For the text data, the training task is purely one of predicting form. For the image data, the training task could be anything, so long as it only involves the images. At test time, we present the model with inputs consisting of an utterance and a photograph, like *How many dogs in the picture are jumping?* or *Kim saw this picture and said "What a cute dog!" What is cute?* and the photos

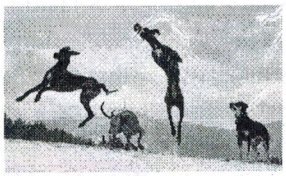
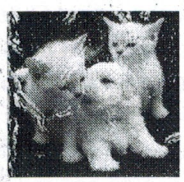

Figure 1: Photo stimuli 1 (L) and 2 (R)

in Figure 1, where the appropriate answers are a number or a region of the photo, respectively.

**Reflections** In both cases, the tests are ridiculous. It seems patently unfair to ask the model to perform them, given what it was trained on. But that is precisely the point we are trying to make: a system that has learned the meaning (semantics) of a programming language knows how to execute code in that language. And a system that has learned the meaning of a human language can do things like answer questions posed in the language about things in the world (or in this case, in pictures).

In other words, what's interesting here is not that the tasks are impossible, but rather what makes them impossible: what's missing from the training data. The form of Java programs, to a system that has not observed the inputs and outputs of these programs, does not include information on how to execute them. Similarly, the form of English sentences, to a system that has not had a chance to acquire the meaning relation $C$ of English, and in the absence of any signal of communicative intent, does not include any information about what language-external entities the speaker might be referring to. Accordingly, a system trained only on the form of Java or English has no way learn their respective meaning relations.

## 6 Human language acquisition

One common reason for believing LMs *might* be learning meaning is the claim that human children can acquire language just by listening to it. This is not supported by scholarly work on language acquisition: rather, we find that human language learning is not only grounded in the physical world around us, but also in interaction with other people in that world. Kids won't pick up a language from passive exposure such as TV or radio: Snow et al. (1976) note in passing that Dutch-speaking kids who watch German TV shows by choice nonetheless don't learn German. Kuhl (2007) shows experimentally that English-learning infants can learn Mandarin phonemic distinctions from brief interactions with a Mandarin-speaking experimenter but not from exposure to Mandarin TV or radio.

Baldwin (1995) and others argue that what is critical for language learning is not just interaction but actually joint attention, i.e. situations where the child and a caregiver are both attending to the same thing and both aware of this fact. This theoretical perspective is substantiated with experimental results showing that toddlers (observed at 15 and 24 months) whose caregivers "follow into" their attention and provide labels for the object of joint attention more have larger vocabularies (Tomasello and Farrar, 1986); that toddlers (18–20 months old) don't pick up labels uttered by someone behind a screen, but do pick up labels uttered by someone performing joint attention with them (Baldwin, 1995); and that at around 10–11 months of age babies pay attention to whether a person's eyes are open or not in terms of whether to follow their gaze, and the degree to which infants in fact follow gaze at 10–11 months while vocalizing themselves predicts vocabulary comprehension 7–8 months later (Brooks and Meltzoff, 2005).[8]

In summary, the process of acquiring a linguistic system, like human communication generally, relies on joint attention and intersubjectivity: the ability to be aware of what another human is attending to and guess what they are intending to communicate. Human children do not learn meaning from form alone and we should not expect machines to do so either.

## 7 Distributional semantics

Distributional semanticists have long been aware that grounding distributional representations in the real world is challenging. The lexical similarity relations learned by distributional models trained on text don't in themselves connect any of those words to the world (Herbelot, 2013; Baroni et al., 2014; Erk, 2016; Emerson, 2020), and the distributions of words may not match the distribution of things in the world (consider *four-legged dogs*).

One approach to providing grounding is to train distributional models on corpora augmented with perceptual data, such as photos (Hossain et al., 2019) or other modalities (Kiela and Clark, 2015; Kiela et al., 2015). Another is to look to interaction data, e.g. a dialogue corpus with success annotations, including low-level success signals such as

[8] These three studies do not name the language that the children were learning. It appears to have been English.

emotional stress (McDuff and Kapoor, 2019) or eye gaze (Koller et al., 2012), which contains a signal about the felicitous uses of forms. The idea that as the learner gets access to more and more information in addition to the text itself, it can learn more and more facets of meaning is worked out in detail by Bisk et al. (2020). We agree that this is an exciting avenue of research.

From this literature we can see that the slogan "meaning is use" (often attributed to Wittgenstein, 1953), refers not to "use" as "distribution in a text corpus" but rather that language is *used* in the real world to convey communicative intents to real people. Speakers distill their past experience of language use into what we call "meaning" here, and produce new attempts at using language based on this; this attempt is successful if the listener correctly deduces the speaker's communicative intent. Thus, standing meanings evolve over time as speakers can different experiences (e.g. McConnell-Ginet, 1984), and a reflection of such change can be observed in their changing textual distribution (e.g. Herbelot et al., 2012; Hamilton et al., 2016).

## 8 On climbing the right hills

What about systems which are trained on a task that is not language modeling — say, semantic parsing, or reading comprehension tests — and that use word embeddings from BERT or some other large LM as one component? Numerous papers over the past couple of years have shown that using such pretrained embeddings can boost the accuracy of the downstream system drastically, even for tasks that are clearly related to meaning.

Our arguments do not apply to such scenarios: reading comprehension datasets include information which goes beyond just form, in that they specify semantic relations between pieces of text, and thus a sufficiently sophisticated neural model *might* learn some aspects of meaning when trained on such datasets. It also is conceivable that whatever information a pretrained LM captures might help the downstream task in learning meaning, without being meaning itself.

Recent research suggests that it is wise to interpret such findings with caution. As noted in §2, both McCoy et al. (2019) and Niven and Kao (2019) found that BERT picked up idiosyncratic patterns in the data for their tasks, and not "meaning". Beyond such diagnostic research on why large pretrained LMs boost such tasks so much, we think there is a more fundamental question to be asked here: Are we climbing the right hill?

### 8.1 Top-down and bottom-up theory-building

There are two different perspectives from which one can look at the progress of a field. Under a *bottom-up* perspective, the efforts of a scientific community are driven by identifying specific research challenges. A scientific result counts as a success if it solves such a specific challenge, at least partially. As long as such successes are frequent and satisfying, there is a general atmosphere of sustained progress. By contrast, under a *top-down* perspective, the focus is on the remote end goal of offering a complete, unified theory for the entire field. This view invites anxiety about the fact that we have not yet fully explained all phenomena and raises the question of whether all of our bottom-up progress leads us in the right direction.

There is no doubt that NLP is currently in the process of rapid hill-climbing. Every year, states of the art across many NLP tasks are being improved significantly — often through the use of better pretrained LMs — and tasks that seemed impossible not long ago are already old news. Thus, everything is going great when we take the bottom-up view. But from a top-down perspective, the question is whether the hill we are climbing so rapidly is the *right* hill. How do we know that incremental progress on today's tasks will take us to our end goal, whether that is "General Linguistic Intelligence" (Yogatama et al., 2019) or a system that passes the Turing test or a system that captures the meaning of English, Arapaho, Thai, or Hausa to a linguist's satisfaction?

It is instructive to look at the past to appreciate this question. Computational linguistics has gone through many fashion cycles over the course of its history. Grammar- and knowledge-based methods gave way to statistical methods, and today most research incorporates neural methods. Researchers of each generation felt like they were solving relevant problems and making constant progress, from a bottom-up perspective. However, eventually serious shortcomings of each paradigm emerged, which could not be tackled satisfactorily with the methods of the day, and these methods were seen as obsolete. This negative judgment — we were climbing a hill, but not the right hill — can only be made from a top-down perspective. We have discussed the question of what is required to

learn meaning in an attempt to bring the top-down perspective into clearer focus.

### 8.2 Hillclimbing diagnostics

We can only definitively tell if we've been climbing the right hill in hindsight, but we propose some best practices for less error-prone mountaineering:

First, above all, cultivate humility towards language and ask top-down questions. Neural methods are not the first bottom-up success in NLP; they will probably not be the last.

Second, be aware of the limitations of tasks: Artificial tasks like bAbI (Weston et al., 2016) can help get a field of research off the ground, but there is no reason to assume that the distribution of language in the test data remotely resembles the distribution of real natural language; thus evaluation results on such tasks must be interpreted very carefully. Similar points can be made about crowdsourced NLI datasets such as SQuAD (Rajpurkar et al., 2016) or SNLI (Bowman et al., 2015), which do not represent questions that any particular person really wanted to ask about a text, but the somewhat unnatural communicative situation of crowdsourcing work. If a system does better on such a task than the inter-annotator agreement,[9] the task probably has statistical artifacts that do not represent meaning. In the vision community, Barbu et al. (2019) offer a novel dataset which explicitly tries to achieve a more realistic distribution of task data; it would be interesting to explore similar ideas for language.

Third, value and support the work of carefully creating new tasks (see also Heinzerling, 2019). For example, the DROP reading comprehension benchmark (Dua et al., 2019) seeks to create more stringent tests of understanding by creating questions that require the system to integrate information from different parts of a paragraph via simple arithmetic or similar operations.[10]

Fourth, evaluate models of meaning across tasks. (Standing) meaning is task-independent, so a system that captures meaning should do well on multiple tasks. Efforts like SuperGLUE (Wang et al., 2019) seem like a good step in this direction.

Finally, perform thorough analysis of both errors and successes. As McCoy et al. (2019) and Niven and Kao (2019) have shown, systems that find success with large pretrained LMs do not necessarily do so because the LMs have learned "meaning". Analyses which start from an attitude of healthy skepticism ("too good to be true") and probing tasks which try to identify what the model actually learned can be good ways to find out whether the system performs well for the right reasons.

[9] https://rajpurkar.github.io/SQuAD-explorer/

[10] See Appendix B for an exploration of what GPT-2 does with arithmetic.

## 9 Some possible counterarguments

In discussing the main thesis of this paper with various colleagues over the past 18 months, we have observed recurring counterarguments. In this section, we address those counterarguments, plus a few more that might arise.

**"But 'meaning' doesn't mean what you say it means."** Defining "meaning" is notoriously hard. For the purposes of this paper, we chose a working definition which is as general as we could make it, capturing the crucial point that meaning is based on the link between linguistic form and something that is not language. "Meaning" cannot simply be the relation between form and some kind of "deep syntax", e.g. semantic dependency graphs (Oepen et al., 2015); like syntax, such representations could perhaps be learned from form alone (He et al., 2018; Hewitt and Manning, 2019). Equating these with meaning ignores a core function of language, which is to convey communicative intents.

**"But meaning could be learned from ...".** As we discussed in §7, if form is augmented with grounding data of some kind, then meaning can conceivably be learned to the extent that the communicative intent is represented in that data.

In addition, certain tasks are designed in a way that specific forms are declared as representing certain semantic relations of interest. Examples of this include NLI datasets (Dagan et al., 2006; Rajpurkar et al., 2016; Ostermann et al., 2019) which pair input/output tuples of linguistic forms with an explicit semantic relation (e.g. text + hypothesis + "entailed"). Similarly, control codes, or tokens like *tl;dr*, have been used to prompt large LMs to perform summarization and other tasks (Radford et al., 2019; Keskar et al., 2019). Here forms are explicitly declared at test time to represent certain semantic relations, which together with the distributional similarity between e.g. *tl;dr* and other phrases such as *in summary*, may be enough to bootstrap a successful neural summarizer. Depending on one's perspective, one may argue that such a system has learned to reliably find instances of the relation without understanding the text; or that

explicitly declaring cues like *entailed* or *tl;dr* as representing certain semantic relations provides a training signal that goes beyond pure form.

Analogously, it has been pointed out to us that the sum of all Java code on Github (cf. § 5) contains unit tests, which specify input-output pairs for Java code. Thus a learner could have access to a weak form of interaction data, from which the meaning of Java could conceivably be learned. This is true, but requires a learner which has been equipped by its human developer with the ability to identify and interpret unit tests. This learner thus has access to partial grounding in addition to the form.

**"But there is *so much* form out there – surely that is enough."** We have argued for the general principle that learning meaning requires more than form. How much form can be observed is not relevant to our point; the octopus can observe A and B for as long as he wants, and the quantity of training data in §5 is not limited.

But given lots of form, could O perhaps learn to keep producing seemingly meaningful responses to A's utterances without learning meaning? The problem is that people constantly generate new communicative intents to talk about their constantly evolving inner and outer worlds, and thus O would need to memorize infinitely many stimulus-response pairs. Such an approach may be an avenue towards high scores in evaluations where perfection is not expected anyway; but it is probably not an avenue towards human-analogous NLU.

**"But aren't neural representations meaning too?"** The internal representations of a neural network have been found to capture certain aspects of meaning, such as semantic similarity (Mikolov et al., 2013; Clark, 2015). As we argued in §4, semantic similarity is only a weak reflection of actual meaning. Neural representations neither qualify as standing meanings ($s$), lacking interpretations, nor as communicative intents ($i$), being insufficient to e.g. correctly build a coconut catapult.

An interesting recent development is the emergence of models for unsupervised machine translation trained only with a language modeling objective on monolingual corpora for the two languages (Lample et al., 2018). If such models were to reach the accuracy of supervised translation models, this would seem contradict our conclusion that meaning cannot be learned from form. A perhaps surprising consequence of our argument would then be that accurate machine translation does not actually require a system to understand the meaning of the source or target language sentence.

**"But BERT improves performance on meaning-related tasks, so it must have learned something about meaning."** It has probably learned *something* about meaning, in the same sense that syntax captures something about meaning and semantic similarity captures something about meaning: a potentially useful, but incomplete, reflection of the actual meaning. McCoy et al. (2019) and Niven and Kao (2019) provide cautionary tales about overestimating what that "something" is purely based on evaluation results on existing tasks. What exactly BERT and its relatives learn about meaning is a very interesting question, and we look forward to further findings from the field of BERTology.

## 10 Conclusion

In this paper, we have argued that in contrast to some current hype, meaning cannot be learned from form alone. This means that even large language models such as BERT do not learn "meaning"; they learn some reflection of meaning into the linguistic form which is very useful in applications.

We have offered some thoughts on how to maintain a healthy, but not exaggerated, optimism with respect to research that builds upon these LMs. In particular, this paper can be seen as a call for precise language use when talking about the success of current models and for humility in dealing with natural language. With this we hope to encourage a top-down perspective on our field which we think will help us select the right hill to climb towards human-analogous NLU.

**Acknowledgments.** This paper benefitted from many inspiring and often spirited discussions. Without implying any agreement with the contents as presented, we thank Sam Bowman, Vera Demberg, Lucia Donatelli, Jason Eisner, Jonas Groschwitz, Kristen Howell, Angie McMillan-Major, Joakim Nivre, Stephan Oepen, Ellie Pavlick, Benjamin Roth, Dan Roth, Asad Sayeed, Hinrich Schütze, Nina Tahmasebi, and Olga Zamaraeva. This paper originated in a Twitter mega-thread that was neatly summarized by Thomas Wolf (2018). We also thank the ACL reviewers and the participants of the Toulouse Workshop on Formal and Distributional Semantics (2015) and *SEM 2016 for their insightful and constructive thoughts.