Distribution Agreement

Signature:

______________________________   ______________
                                                                        Date

The Generative Nature of Commonsense Knowledge:
Insights from Machine Learning

By

Jacquelyn M Ellison
Master of Science

Psychology

____
Phillip Wolff, Ph.D.
Advisor

____
Robyn Fivush, Ph.D.
Committee Member

____
Harold Gouzoules, Ph.D.
Committee Member

Accepted:

____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____
Date

The Generative Nature of Commonsense Knowledge: Insights from Machine Learning

By

Jacquelyn M Ellison
B.A., University of Colorado Boulder, 2015
M.A., University of Colorado Boulder, 2015

Advisor: Phillip Wolff, PhD

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Psychology
2020

Abstract

The Generative Nature of Commonsense Knowledge: Insights from Machine Learning
By Jacquelyn Ellison

The study of commonsense has received little attention for lack of accounts of how it might be represented. Recent advances in machine learning are resulting in rich knowledge bases that (unwittingly) might offer some insight into this elusive phenomenon. This paper assesses one pre-computed model, RoBERTa, for suitability as a working model of human commonsense knowledge by testing it against variation in human agreement. We examine the contribution of statistical and structural properties of language to this performance, including frequency and cooccurrence-based representations, as well as part of speech and syntactic structure. We conclude that RoBERTa is a suitable model for language prediction: the model's predictions closely reflected human agreement and cannot be explained by simple linguistic features. In investigating the range of possible responses to a particular context, we find that these responses illustrate the impact of categorical organization on precise context sensitivity and conclude that this demonstrates the hallmarks of commonsense knowledge. After exploring the contribution of the static component of RoBERTa's knowledge, the main finding of this paper is that the knowledge base that directly facilitates both human agreement and the model's measure of fit is by its very nature generative, and only truly exists in representation as it is applied. This paper discusses the role of implicit learning and predictive processing as potential frameworks within which to substantiate this meta-theoretic observation.

The Generative Nature of Commonsense Knowledge: Insights from Machine Learning

By

Jacquelyn M Ellison
B.A., University of Colorado Boulder, 2015
M.A., University of Colorado Boulder, 2015

Advisor: Phillip Wolff, PhD

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Psychology
2020

TABLE OF CONTENTS

Introduction

Knowing how the world works is critical to navigating it successfully, but the origins of this knowledge are poorly understood. Regardless of whether they have witnessed such an event, humans know that a boulder can smash an umbrella, but a penny cannot, and that a hundred peas can fit into a shoe, but not into a spoon. This common knowledge underpins everyday tasks from language comprehension to decision making and is rarely acquired through explicit instruction (Wagner & Sternberg, 1986; Sternberg & Caruso, 1985). It displays precise sensitivity to novel contexts while maintaining consistency across individuals and events, a conundrum for theories of memory that reference accessible but static store of knowledge (Moscovici & Hewstone 1983). Much like autobiographical memory (and episodic memory more broadly), it must be produced spontaneously in response to unique circumstances. Common knowledge is, therefore, knowledge that exists in its application, as when solving problems or making predictions.

Understanding the intricacies of such a system has been a core concern in artificial intelligence but recently has seen limited interest in other areas of cognitive science. For a novel insight into this problem, we turn to an unlikely source; a new class of language models that ostensibly do not seek to model common sense nevertheless demonstrate behavior that suggests knowledge of the world. These models bring new relevance to old findings and suggest a novel approach for understanding what it means for knowledge to be inherently applied.

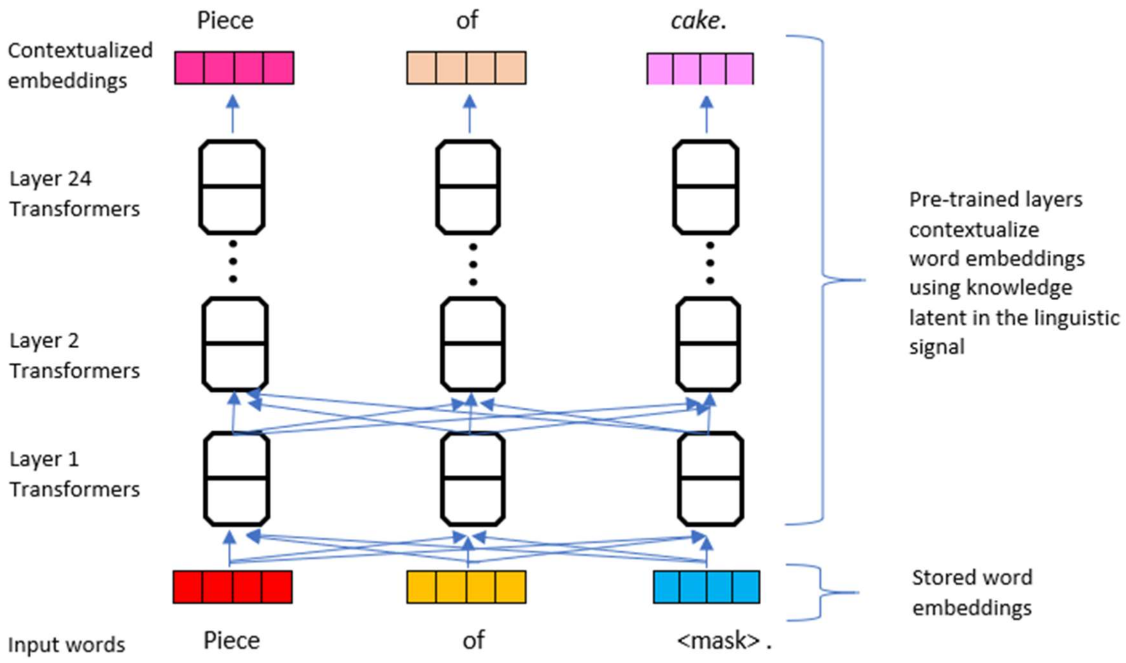
A growing body of evidence suggests that actively predicting language as we hear it drives the acquisition of knowledge that is implicit in the linguistic signal (Köster et al., 2020). A concrete and robust example of how knowledge can be inferred from language is the perceptual learning of phonetic categories. By the age of 6 months, infants can discriminate sounds that contrast in their native language better than those that do not (Kuhl et al., 1992; Werker & Tees, 1984). Findings like those from Maye et al. (2002) suggest that this change in phonemic discrimination is driven by the pattern of statistical regularities in the sounds of adult speech (Werker et al., 2012 for review), which in turn facilitates their prediction in context (Zettersten, 2019 for overview). In other words, this is knowledge that is gained through its application in prediction. Prediction-based sensitivity to the information latent in the signal appears to be a general principle of learning from language, encompassing word learning (Saffran et al., 1996; Estes et al., 2007), simple pattern learning (Marcus et al., 1999), and complex phrase-structure learning (Thompson & Newport, 2007). Critically, it's not only the forms of language that can be discovered from statistical regularities in experience; word meanings can be inferred by 12-month-old infants by attending across scenes to cues that would be ambiguous on their own (Smith & Yu, 2008). Using language as the only source of input, both word meanings (Landauer & Dumais, 1997; Mikolov et al., 2013) and more highly abstract forms of knowledge, like social biases (Caliskan et al., 2017), have been acquired computationally from large-scale text corpora. Sternberg et al. (1995) even argue that, much like word meanings and social biases, the commonsense knowledge that serves everyday practical intelligence can be learned implicitly.

Despite the substantial and growing support for this principle, current computational efforts seek to model commonsense knowledge with static databases. Their approach is that of the encyclopedist; both the Cyc project and MIT's Open Mind Common Sense project, as well as several others that work in a similar spirit, rely on thousands of hours of manually-generated facts (generated by the researchers at Cyc, or through crowd-sourcing at Open Mind), curated and structured with varying degrees of automation (Lenat et al., 1990; Singh et al., 2002; Bollacker et al., 2007; Cambria et al., 2014). The goal of these projects is to capture the summed mass of human knowledge for the purpose of improving efforts in artificial intelligence; however, failing to recognize that commonsense knowledge is itself a form of applied intelligence results in a brittle database. Though well-structured and outfitted with many tidy logical operators for inference engineering, both projects are unable to adapt to new information without man-hours of input and are fully reliant on finite sets of relational structures.

Where these models fail, however, a new class of transformer-based language models succeeds. This new class of models appears to acquire implicit knowledge from language and adapts it flexibly to new applications. One such model is RoBERTa, a pretraining variant of the BERT-Large architecture that achieves human-like performance on a wide range of Natural Language Processing (NLP) tasks (Liu et al., 2019; Devlin et al., 2018). During training, RoBERTa learns both semantic representations for words and the relational structures that bind them into coherent sentences. The representations themselves are word embeddings, arrays of numbers that specify a location in semantic space. Therefore, these embeddings allow semantic relations to be encoded as spatial relations. In addition to these embeddings, RoBERTa learns how to assemble them meaningfully into sentences, through a generative architecture that itself is trained along with the embeddings. As shown in Figure 1, this model learns through a process of text prediction: given a sentence in which a selection of words have been masked out, RoBERTa attempts to fill in the blanks, and learns to the extent that this attempt is in error. In other words, this training task only requires the model to improve in resolving the ambiguity in a noisy signal. This is accomplished both by improving the prediction process (the architecture) and improving the representations that serve as input for that process (the embeddings).

Figure 1

Diagram of Transformer-Based Architecture and Text Prediction



Note: In training and in text prediction, word embeddings are adjusted to reflect their context. This contextualization is the result of the attention weights produced by each transformer (blue arrows). These weights are fully interconnected, so that the entire context is applied to each word at every layer. This allows a meaningless token, *<mask>* to take on the meaning suggested by its context, here, *cake*. This diagram is adapted from Rezaii et al. (2020).

RoBERTa's architecture consists of 24 layers of transformer encoders, amounting to 345 million parameters, and was pre-trained on about 160GB of text from 4 corpora. Once trained, the model can be fine-tuned by adding a feed-forward layer and training a little longer for specific tasks. On the wide variety of language understanding tasks represented in the GLUE benchmark, RoBERTa achieves near-human levels of performance (Wang et al., 2018). This diverse set of tasks includes detecting whether a sentence is positively or negatively valenced, determining the logical relation between a pair of sentences, answering questions and paraphrasing text, among others. These tasks are designed to be easy for humans and very difficult for machines. In humans, they require knowledge of linguistic structure, word meanings as well as sentence-level semantics, causality, entailment and other types of abstract relations. In other words, these tasks seem to rely both on linguistic knowledge and knowledge that is not specific to language. Seeking to understand both RoBERTa's performance and its knowledgebase may yield powerful insights into the origins of common-sense knowledge.

The most basic task in RoBERTa's repertoire, one that requires no fine-tuning, is word prediction. In response to a sentence with a word masked out, like *The man <mask> a gallon of milk*, RoBERTa returns the vocabulary, rank-ordered according to fit, as illustrated in Table 1. The values accompanying the words are logits, the model's measures of fit, where a higher number indicates a better fit.

Table 1.

RoBERTa's predictions for the missing word, <mask>

	Sentence 1
Input	The man <mask> a gallon of milk.
Response	*bought* (57.18), *ordered* (56.45), *buys* (56.06), *had* (55.91) …

Critically, this performance is sensitive to context. In Table 2, note that not only did the responses shift in response to the subtle change in context (a single word), the fit also improved relative to the example in Table 1.

Table 2.

RoBERTa's predictions for the missing word given more specific context clues

	Sentence 1
Input	The man went to the store. He <mask> a gallon of milk.
Response	*bought* (62.31), *purchased* (59.87), *got* (59.61), *buy* (58.22) …
	Sentence 2
Input	The man went to the restaurant. He <mask> a gallon of milk.
Response	*ordered* (60.73), *bought* (59.79), *got* (58.66), *purchased* (58.21) …

This sensitivity to context is both flexible and precise, as illustrated by the ability to select, out of a vocabulary of 54,597 tokens, not only the correct part of speech, and an appropriate activity to do with milk, but also the right verb for the location specified in the first sentence. In an effort to understand this performance and the knowledge it draws upon, we designed a series of studies to investigate the extent to which this behavior is attributable to complex knowledge structures and the extent to which this knowledge is generated on-line in response to specific contexts.

Study 1: The Contributions of Linguistic Structure and Distribution

Our first concern was to understand how human-like this model's behavior is. While its overall performance is very impressive on tasks at which humans excel, it remains unknown whether this performance varies with human performance. Are humans and RoBERTa using the same cues to solve their problems? Are they drawing on the same knowledge? If so, this would open the possibility of exploring whether and how commonsense knowledge is created through language understanding.

To begin exploring this question, we first asked about the features that determine the predictability of a sentence. If RoBERTa is in agreement with humans about what sentences are predictable, then this would suggest that it may be a reasonable model with which to explore this type of intelligence. However, it will be important to know whether this agreement is due to interesting reasons, such as the presence of generative knowledge in RoBERTa, or less interesting reasons, such as low-level statistical associations. Since this model learns through the statistical regularities in the linguistic signal, there are several kinds of statistical regularities RoBERTa might use to generate responses. Perhaps more frequent words are simply easier to fit into any given context, and therefore easier to predict. Alternatively, it might be slightly more complex, and prediction is driven by the regular coocurrence between a word and its context. However, it could be that sentences that are distributionally coherent pull from the same semantic 'neighborhood,' and this coherence makes it easier to predict the missing word. Maybe it's just that longer sentences contain more information, and this is what makes the difference.

On the other hand, this predictive ability may reflect more complex knowledge of language. Perhaps the model knows part-of-speech, which is systematically distributed through language use. Nouns and Verbs are more concrete and specific to particular contexts, so perhaps they would be easier (or harder, for that matter) to predict than prepositions. Alternatively, maybe RoBERTa really is a sophisticated language user, in which case it may be that the syntactic structure of the sentence plays a role in how predictable the missing word is.

Study 1 tests the extent to which these linguistic features explain the variation in RoBERTa's performance and in the level of agreement between humans. We make use of the well-known cloze procedure, which has been in use since 1953, not only in empirical research, but in classrooms as an assessment of reading and general native and second language ability (Aitken, 1977; Taylor, 1953). This procedure consists of predicting the final word of a sentence, which has been obscured or deleted. The proportion of respondents who complete the sentence with the same word (the cloze probability for that sentence), is a reliable measure of how predictable that response is, or from another perspective, how strongly that sentence constrains the range of possible responses. During training, RoBERTa receives feedback on cloze completions, a procedure known to improve reading comprehension in humans (Schneyer, 1965; Apsari, 2016).

Methods

Participants Subjects were individuals who participated in the rating tasks reported in Peelle et al. (2020), *N* ≥ 100 MTurkers for each sentence.

Materials The stimuli consisted of 3,085 sentences from Peelle et al. (2020) with cloze norms for human completions. Of these, 242 were excluded to accommodate RoBERTa's whole-word vocabulary of 33,135 words, resulting in 2,843 sentences spanning a wide range of cloze probability. For all sentences, we used the most frequent human response as the expected response and its cloze probability as a measure of sensitivity. This cloze probability is the proportion of human participants who completed the sentence with the most frequent response. Table 3 shows an example of a high-cloze sentence (above the 90th percentile of cloze probabilities represented in our stimuli) and a low-cloze sentence (below the 10th percentile) from this set. In Table 3, each human response is accompanied by the probability that that response was given; the expected response for each sentence is in bold.

Table 3

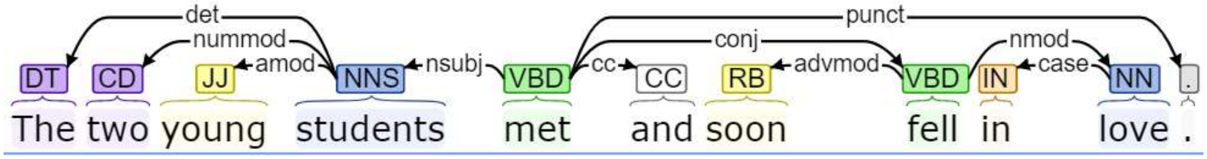
Samples of High- and Low- Cloze Norms

Cloze Value	Example Sentence
High Cloze	The old house will be torn <mask>. <mask> → ***down*** *(0.93), apart (0.02)*
Low Cloze	Helen reached up to dust the <mask>. <mask> → ***shelf*** *(0.37), cabinet (0.18), mantle (0.15), lamp (0.06), counter (0.05), ceiling (0.02), chair (0.02), closet (0.02), chandelier (0.02), table (0.02)*

We also computed several linguistic measures on the sentences themselves. Each sentence was parsed using the StanfordNLP neural pipeline, which gave us the part-of-speech for each expected response, as well as the dependency relation linking the expected response to the sentence (Manning et al., 2014; Toutanova, 2003; Chen & Manning, 2014). As shown in Figure 2, dependency relations indicate a word's role in a sentence. This dependency information was used to compute a word's grammatical centrality, that is, the inverse distance from the main verb. Centrality is measured relative to the main verb because this is believed to serve as the syntactic 'core' of a sentence (Carnie, 2012). Distance from the main verb was determined by counting the number of relations between the expected response and the main verb in the dependency graph. An additional measure was sentence length, which we determined by counting the number of words in a sentence.

Figure 2

Dependency parse and centrality metric



Note: A Dependency parse shows the parts-of-speech of the words in the sentence and the grammatical relations (e.g., subject, direct object) between words. The expected response for this sentence is *love*, which is connected to the main verb by two relations, first, as a noun modifier to the verb, *fell*, and second, through the conjunction relation between *fell* and the main verb, *met*. The distance from the verb is 3 (a distance of 1 would be the verb itself), so the centrality measure is 0.33.

We also made use of a number of pre-computed linguistic statistics, the log-normalized word frequency for each response from the English Lexicon Project (Balota et al., 2007), and cooccurrence frequencies from the collocates database of the Corpus of Contemporary American English (COCA) (Davies, 2008). The collocates database was used to develop two measures of predictability for each cloze sentence. The average coocurrence of all word combinations in the sentence excluding the <mask> served as a measure of distributional coherence for each sentence, while the average frequency with which each word in the sentence coocurs with the expected response served as a measure of distributional predictiveness.

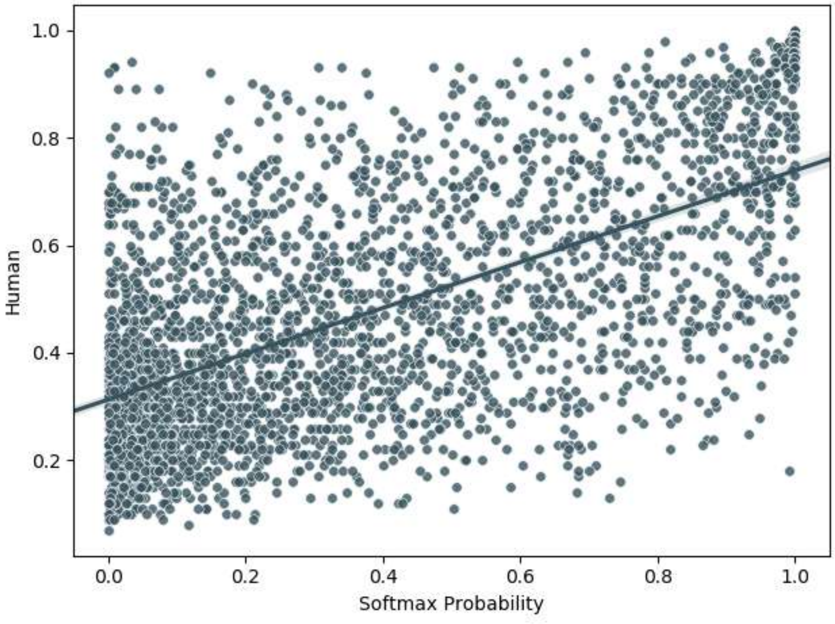
For a measure of RoBERTa's performance, we used the measure of fit that is attached to each of its responses. This measure of fit is the logit, or the log-odds of the response probability.

Procedure We used the Huggingface version of RoBERTa, available on github (https://github.com/huggingface/transformers). Programs were written to retrieve the logits for each word in the lexicon, which were assigned by RoBERTa in response to the sentence context. The logits were rank-ordered to determine the words representing RoBERTa's guesses about the words that best filled the missing word in each test sentence.

Results RoBERTa's response mirrored the cloze probabilities of humans well. Based on only RoBERTa's first choice, the proportion of expected responses that were among Roberta's best-fit for the context was 45.35% and 60.48% including the second choice. This level of performance compared favorably with that of humans. The average percent agreement of a human participant other humans, that is, with the modal response of the group was 48.55% (*SD* = 23.20). For more direct comparison with human cloze probabilities, a softmax function was computed over the logit values, re-scaling them to within (0, 1). As shown in Figure 3, these values strongly correlated with human cloze probability, *r*(2842) = 0.60, *p* < 0.0001. Like humans, RoBERTa is sensitive to how strongly a sentence context constrains the possible completions, and, even more compelling, this model agrees with humans about which sentences exert a stronger constraint on the range of possible completions.

Figure 3

RoBERTa' Probability of Predicting the Expected Response by Human Cloze Probability

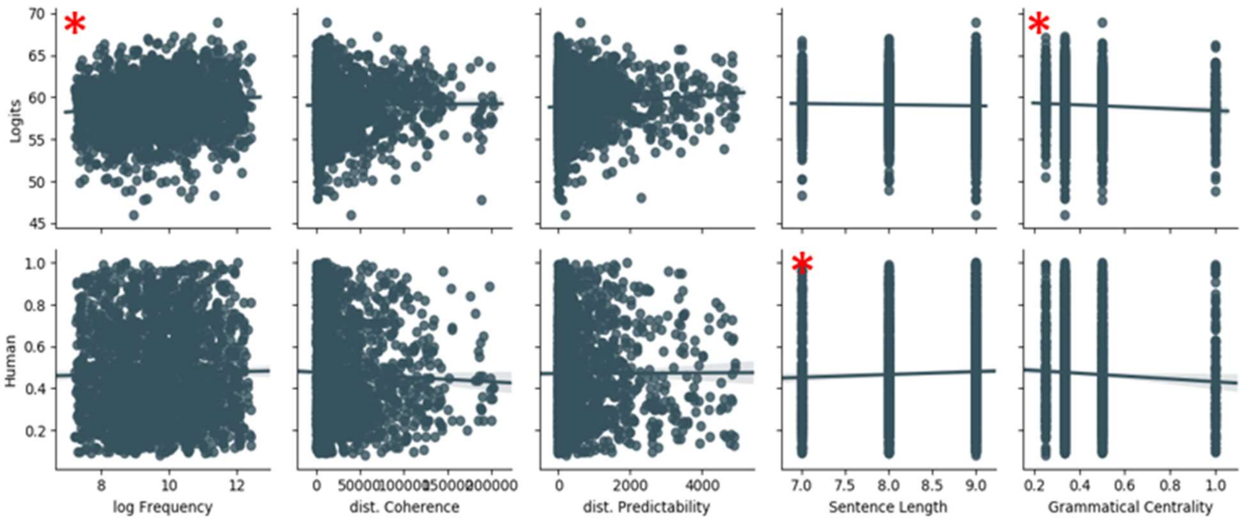


Note: The probability assigned by RoBERTa to the expected response is shown on the y-axis, plotted against human cloze probability, on the x-axis. These measures are strongly correlated with *r*(2842)=0.60.

Because the softmax function changes the distribution of values (high values get squashed to the ceiling, while low values get squashed to the floor), the untransformed logits will be used for all analyses that don't directly compare human and model performance. The following analyses investigate some possible explanations for the high agreement between RoBERTa and humans. In particular, it investigates whether the agreement might be explained by relatively uninteresting, simple linguistic features such as word frequency, grammatical centrality, part-of-speech (POS), sentence length, and distributional coherence. Regressing RoBERTa's predicted logits on all of these variables indicated that low-level linguistic features could explain a portion of the variance, $F(10, 2830) = 18.53$, $p > .0001$, but not much, as indicated by a small adj. R^2 of 0.058. In the full model, the corpus frequency of the expected response ($\beta = 0.36$, $p > .0001$), and the grammatical centrality ($\beta = -0.94$, $p = .005$), as well as three parts of speech (nouns ($\beta = -1.36$, $p = .044$), adjectives ($\beta = -2.15$, $p = .002$), and verbs ($\beta = -1.41$, $p = .038$)) were significant predictors. A second regression was run using the same independent variables to predict human cloze probability. As in the analysis with RoBERTa, these low-level language measures explained only a small proportion of the variance, $F(10, 2830) = 4.18$, $p < .0001$ with an adj. R^2 of 0.011. In contrast to model fit, human agreement was not predicted by either the frequency of the response or its grammatical centrality, however sentence length ($\beta = 0.014$, $p = .007$) as well as the indicators for nouns ($\beta = -0.12$, $p = .042$) and adjectives ($\beta = -0.13$, $p = .009$) contributed to the fit of the model. The scatter plots in Figure 4 allow us to compare the results from humans and RoBERTa with respect to these linguistic variables, excluding part-of-speech. As can be seen visually, the distributional coherence, distributional predictability, log frequency, sentence length and grammatical centrality accounted for little of the variance in the choices made by RoBERTa or Humans.

Figure 4

Scatter Plots of Human Agreement and Model Fit with Respect to Five Linguistic Measures



Note: None of the linguistic structural or distributional measures we looked had had a strong relationship with predictability, either for the model's fit (top) or for human agreement (bottom). From left to right, these measures are the log frequency of the response, the distributional coherence of the sentence, the distributional predictability of the response, the sentence length, and grammatical centrality of the response. *indicates a significant result

Discussion RoBERTa and humans are in agreement about what a predictable sentence looks like, and predictability seems to have very little to do with the number of words in a sentence, the structure of that sentence (as current theory knows it), the basic statistical properties of the words in the sentence, or the part-of-speech of the word you're trying to predict. On the basis of these measures, it there is little evidence that RoBERTa's performance can be explained away by relatively simple linguistic factors.

Study 2. Context-sensitivity and the structure of the knowledge base

RoBERTa's response to a missing word is a probability distribution over every token in the vocabulary. We took advantage of this variation of fit within a response to understand whether RoBERTa's performance indicates the kind of well-structured knowledge base necessary for using commonsense knowledge rapidly and flexibly. In an investigation of RoBERTa's parent model, BERT, Ettinger (2019) found no evidence that BERT possessed such a knowledge base. Following her lead, we investigated RoBERTa's performance on a carefully-constructed set of sentence pairs from Federmeier and Kutas (1999) that were designed for an ERP study testing whether categorical structure influences the use of context in online language processing. We only used sentences from Federmeier and Kutas (1999) whose responses were included in RoBERTa's fixed vocabulary of about 33,135 words. Table 4 lists some of the stimuli included in the dataset. Each sentence-pair is shown with three possible completions, crossing levels of expectedness. The 'expected' response is the one humans prefer (in their study, Federmeier and Kutas found an average cloze probability of 0.74, i.e., 74% of participants spontaneously produced the same word in response to the sentence cue). The 'within-category' ('within') responses were implausible completions from within the same semantic category as the expected response, and the 'between-category' ('between') responses are both implausible and from a different but related category.

Table 4

Sample Sentences from Federmeier and Kutas (1999)

	Sentence Cue	**Expected** | within | between
1	Getting both himself and his car to work on the neighboring island was time-consuming. Every morning he drove for a few minutes and then boarded the <mask>.	**Ferry** | gondola | plane
2	The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a <mask>.	**Shovel** | rake | saw

3	He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that <mask>.	**Lipstick** | mascara | earring

As Ettinger (2019) points out, successful completion of the second sentence requires use of commonsense knowledge and complex inferencing of what the content of the first sentence might entail for the situation described in the second sentence. For instance, in sentence pair 1 from Table 4, the second sentence requires a form of transportation, but all of the information needed to decide between the expected response, *ferry*, and any other transport is given in the first sentence. Commonsense knowledge gives the understanding that travel to an island requires either a watercraft or an aircraft, the understanding that this form of transportation would need to run on a regular schedule and would most likely be open to public use, and the knowledge that it would need to be equipped for the transportation of a car. Pragmatic reasoning, meanwhile, brings the understanding that the first sentence is the statement of an unexplained summary perception, and the second sentence will likely elaborate, using the information from the first sentence. This set of stimuli is therefore well-suited to understanding the extent to which RoBERTa possess commonsense knowledge and the extent to which this knowledge is organized into human-like categories.

Methods

Participants The participants in Federmeier and Kutas (1999) were 170 UCSD undergraduate volunteers.

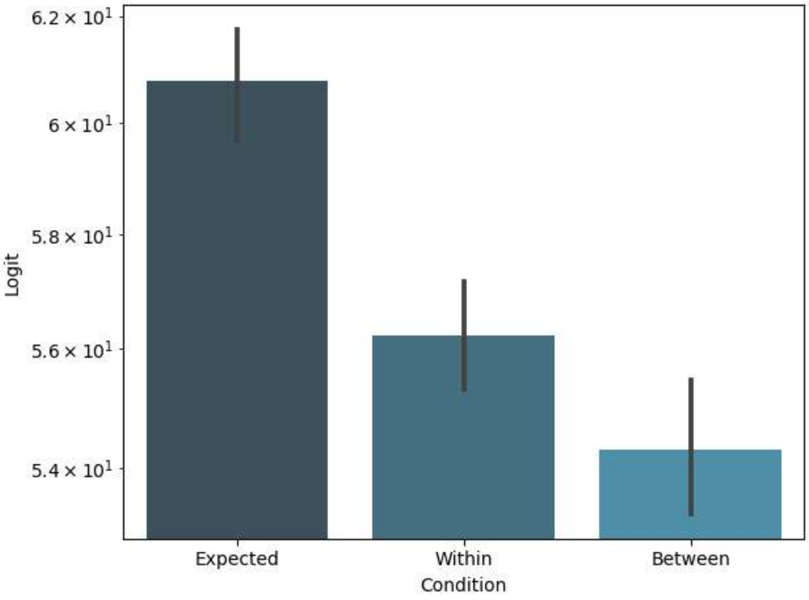
Materials Sentence pairs were 34 of the 40 made available by Federmeier and Kutas (1999). Six of the sentence pairs could not be used because they contained words not present in RoBERTa's whole-word lexicon. Example sentence pairs are shown in Table 4.

Procedure This study used the same procedure as Study 1 and varied only in the stimuli used.

Results RoBERTa was moderately successful in completing the sentence pairs. Federmeier and Kutas (1999) found a mean cloze probability for the expected tokens of 0.74, indicating that the degree to which these sentences semantically constrain their responses is moderate, and calling for a more lenient criterion. RoBERTa's top-ranked response matched the expected response 58.82% of the time. When RoBERTa's prediction was extended to include the top two responses, the modal response was selected 88.23% of the time, which is higher the human average of 74% in Federmeier and Kutas (1999). Interestingly, RoBERTa's confidence in different categories of responses—as reflected in logits--mirrored that of humans. As shown in Figure 4, the average logit for expected ($M = 46.18$; $SD = 3.17$), within-category ($M = 42.74$; $SD = 2.94$) and between-category ($M = 41.29$; $SD = 3.40$) responses, decreased with semantic distance from the expected responses, $F(2, 33) = 37.25$, $p < 0.001$, $\eta^2 = 0.429$, with poorest fit reflected in the between-category violations and the strongest fit to the expected exemplars. Two post-hoc comparisons were conducted, assuming a Bonferroni-corrected $\alpha = 0.025$. The fit for expected responses was higher than within-category responses ($p < .001$), which in turn were higher than between-category responses ($p = .015$).

Figure 5

Logit by Response Type



Note: Mean logits generated by RoBERTa for the Expected words, Within-category words, and Between-category words. Error bars show 95% confidence intervals. The logits on they y-axis are log-scaled.

Discussion This experiment sought to evaluate whether RoBERTa's knowledge base is structured into categories, like human semantic memory, and whether the model's notable sensitivity to context is colored by this structure. Much like the graded response of the N400 ERP component found by Federmeier and Kutas (1999), the graded response we observed in RoBERTa's measure of fit suggests that while the immediate context contains enough information to decide between semantically similar (with similar patterns of co-occurrence) category members, this model's prediction process is also sensitive to the categorical relationship between potential completions. In other words, context matters, but so does semantic knowledge. As argued by Ettinger (2019), the cloze task from Federmeier and Kutas (1999) appear to require use of commonsense knowledge. RoBERTa's success in predicting the expected response with these materials suggests that it may possess a limited degree of human-like commonsense.

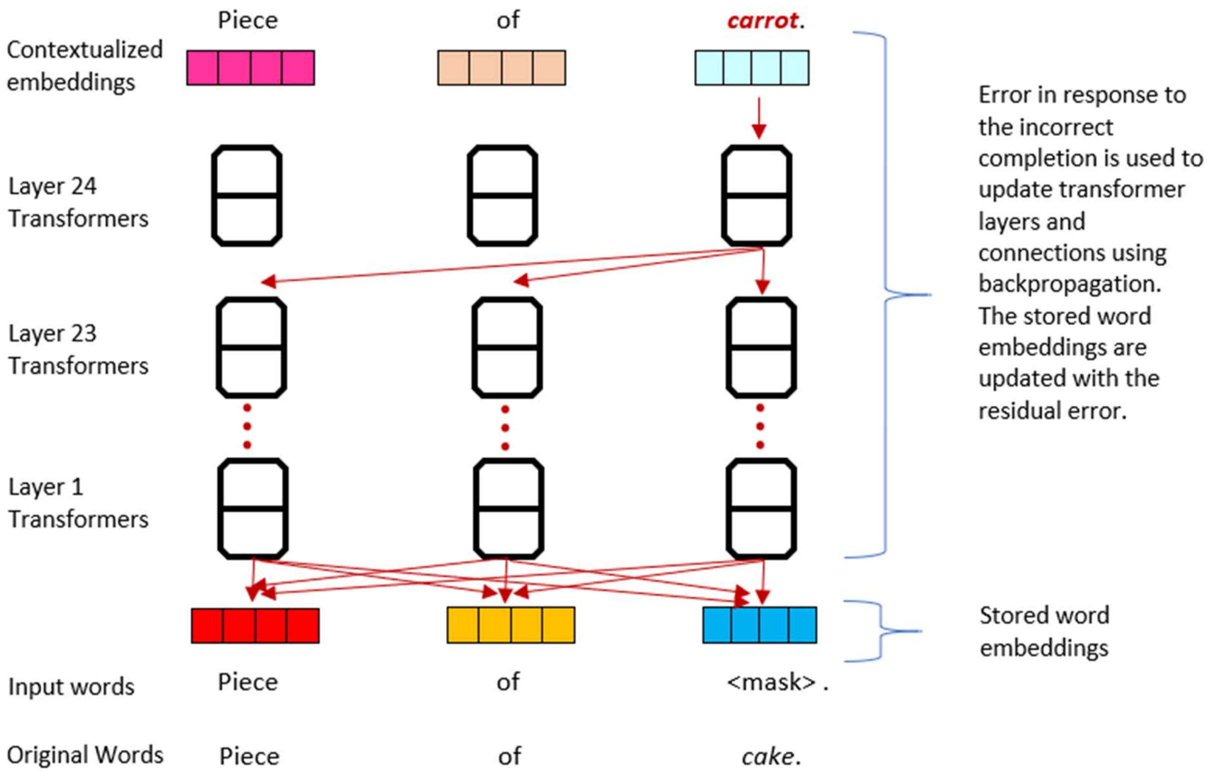
Study 3: Understanding RoBERTa's Knowledge Base

Knowledge of linguistic structure and distribution is insufficient to explain RoBERTa's human-like performance, but what knowledge would be sufficient? When we took a closer look at RoBERTa's knowledge base, we found that it is in fact split. Part of what is updated during training is a set of word embeddings, analogous to those produced by GloVe or Word2Vec (Pennington et al., 2014; Mikolov et al., 2013). These word embeddings are vector representations of words, which start off as random lists of numbers, but are trained through exposure to text. In Word2Vec, word embeddings are generated by training a 3-layer neural network to predict the context they most likely occurred in. The error that is generated in response to an incorrect prediction is used to update both the weights in the architecture itself and the embeddings. Since the architecture is small, there are few weights to update; as a result, the embeddings themselves store a large proportion of the information needed to make accurate predictions. Similar embeddings have demonstrated remarkable performance in text representation but are static, that is, their word representations do not change across contexts. It is therefore believed that language understanding tasks, like recognizing the sentiment of a paragraph or identifying the logical relation between a pair of sentences, are beyond their capacity.

As with Word2Vec, RoBERTa generates word embeddings. However, the process of relating these words to one another in meaningful contexts is a property of the architecture itself. This architecture is also updated during training, a process through which it learns to make predictions with uncanny precision. In fact, the embeddings are only updated in response to the residual error that remains after distributing it throughout the architecture's 24 layers, each equipped with sub-structures (transformers) that are themselves trained, as illustrated in Figure 6. As a result, like other transformer-based models, RoBERTa's knowledge is housed in both the static word embeddings, and in the machinery that processes them.

Figure 6

Error Updating During Training in Transformer-Based Models


Piece
of
carrot.
Contextualized embeddings
Layer 24 Transformers
Layer 23 Transformers
Layer 1 Transformers
Error in response to the incorrect completion is used to update transformer layers and connections using backpropagation. The stored word embeddings are updated with the residual error.
Stored word embeddings
Input words
Piece
of
<mask> .
Original Words
Piece
of
cake.

Note: Colored blocks represent the embedded vector representation for the word immediately above or below it. In training, when a model like RoBERTa produces a word for the *<mask>* that is different from the one in the original sentence, an error is generated. Using backpropagation, this error (red arrows) is distributed across the 24 layers and back to the stored word embeddings, resulting in a relatively small proportion of error being used to update the embeddings themselves. The result is a flexible knowledge-base largely stored in the architecture itself, while the static embeddings Diagrams adapted from Rezaii et al. (2020).

One key feature of common knowledge, according to Moscovici and Hewstone (1983) is that it is inherently applied. Though this may be enigmatic, understanding the knowledge underlying RoBERTa’s performance could provide the insight needed to explore how this could be the case. In pursuing this question, our first concern is understanding whether the RoBERTa’s generative process is necessary for completing sentences in such human-like ways. If the static embeddings are able to describe human variation as well as RoBERTa's full model, this would suggest that the machinery that carries out the prediction is no more than a highly complex and sophisticated symbol-shuffler. However, if the opposite is true; that is, if this static knowledge describes very little of the variance in the model's performance, this would suggest that the generative process itself contains the bulk of the knowledge that determines whether a given sentence completion will be predicted. In other words, we would have evidence that the completion process depends on generative processing.

Such a conclusion assumes, however, that the lower performance of static word embeddings is not simply due to the embeddings being poorly trained. To confirm that the static word embeddings are contentful, an additional analysis was conducted. Specifically, we investigated whether the static word embeddings could perform simple, context-free, similarity judgments using the WordSim353 dataset, which has been used extensively in evaluating embeddings (eg, Baroni et al., 2014; Levy & Goldberg, 2014). This dataset consists of 353 word-pairs and human judgments of the relatedness between those words, on a ten-point scale. The word embeddings are considered contentful to the extent that the similarity judgments they produce agree with humans.

The current study used two kinds of static word embeddings. RoBERTa's embeddings represent the static knowledge base, before any generative processing. The second kind of static word embedding were those generated by an algorithm known as GloVE (Pennington et al., 2014). The GloVe embeddings contain representations for nearly all of the words in the English language. Moreover, these embeddings have been well-validated on many linguistic tasks and have been used widely to achieve various natural language processing goals (e.g., Gupta et al., 2018; Dhingra et al., 2017).

It is worth noting that while the term *similarity* is used most frequently to describe the "closeness" of word vectors in vector-space, a more accurate term is *relatedness*, since the semantic notion of *similarity* maps poorly to the relationship between *cup* and *saucer*, though these regularly occur in close proximity in vector sets. In addition to simple relatedness, it has been shown that vector arithmetic can reveal more specific semantic relationships (Mikolov et al., 2013; Bolukbasi et al., 2017). For instance, an analogy of the form *man:king :: woman:X* can be successfully solved by taking $\overrightarrow{king} - \overrightarrow{man}$, and adding the difference to $\overrightarrow{woman}$. This new vector will be approximately equal to $\overrightarrow{queen}$. Similarly, Lev et al. (2015) attained performance comparable to then-state-of-the-art in generic text representation by using an average of the word vectors in a sentence as a semantic representation of the sentence. Using similar methods, we can approximate the cloze procedure used in Study 1, using only the relatedness captured by the embeddings.

Methods

Participants Subjects were individuals who participated in the rating tasks reported in Agirre, et al. (2009), N = 23, and in Peelle et al. (2020), $N \geq 100$ MTurkers for each sentence.

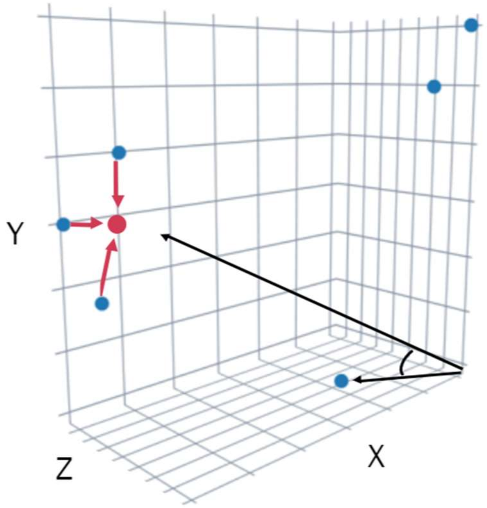
Materials The stimuli set we used in this study is the same 3,049 cloze sentences used in study 1. In addition to RoBERTa's embeddings, we tested our procedure on the widely used pre-trained GloVe embeddings from Pennington et al. (2014). Both sets of embeddings were first evaluated using the WordSim353 dataset from Agirre, et al. (2009).

Procedure To evaluate the quality of the embeddings, we computed the cosine similarity between the corresponding vectors for each word-pair in the WordSim353 dataset and compared the result to the human similarity judgments.

Additionally, we approximated the cloze procedure using each set of word embeddings. We accomplished this by taking the average over all the word embeddings in the sentence to produce a new sentence embedding. We next identified the word embedding associated with the expected completion for that sentence, and computed the cosine-similarity between the sentence embedding and the expected word as illustrated in Figure 6. This cloze cosine was used as the measure of sensitivity for each set of embeddings, and was compared to the logits generated by the full model in study 1, and the cloze probabilities for human responses, as reported in Peelle et al., 2020.

Figure 6

Illustration of Embedding Cloze Procedure



Note: The three blue points nearest to the labeled *Y* axis represent the vectors for the words in a given sentence. These were averaged (illustrated by red arrows) to produce a new point, shown in red. From this new point, similarity to the response, indicated by the point nearest to the *X* axis was measured by computing the cosine of the angle between the two, with its vertex at the origin.

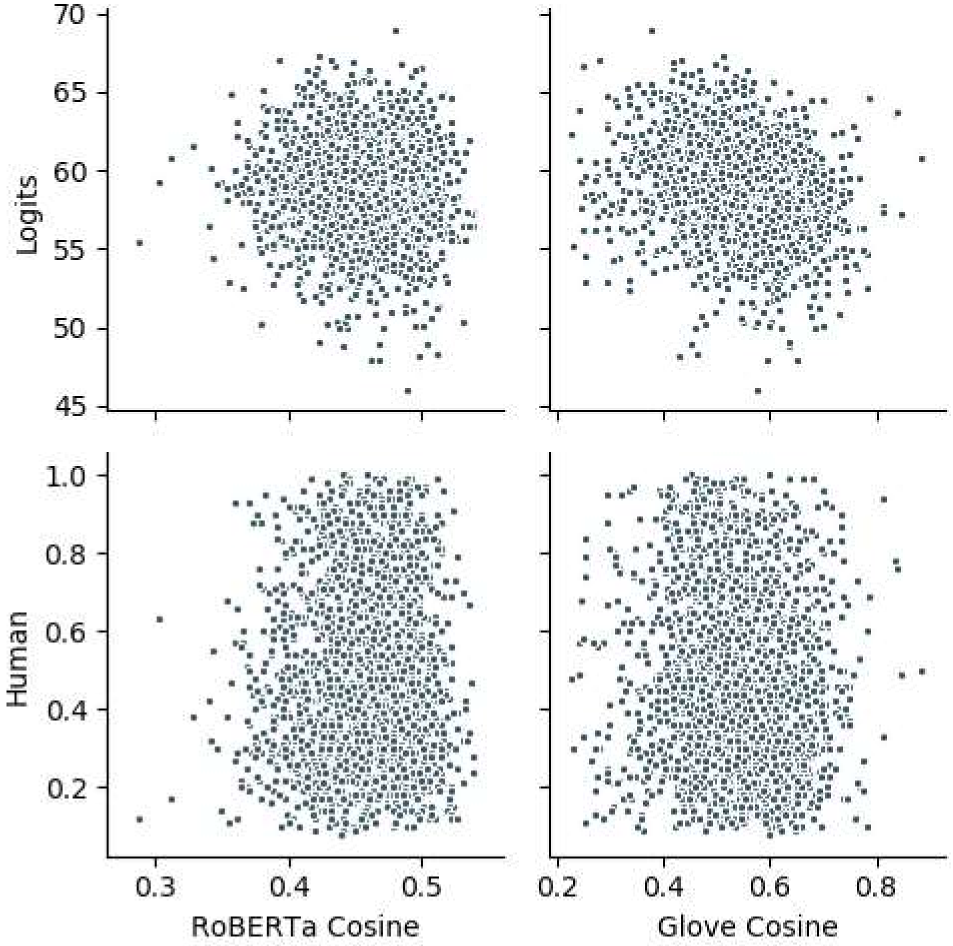
Results

Cloze Task As shown in Figure 7, RoBERTa's cloze cosines did not correlate with human cloze probabilities, *r*(2842) = 0.003, *p* = .843, suggesting that they are a poor reflection of the information humans use to complete these sentences. Though the correlation between RoBERTa's cloze cosines and logits was significant, it was also small, *r*(2842) = -0.070, *p* = 0.0002, indicating that these embeddings do not show the same sensitivity to context that is shown by the full model. Interestingly, the cloze cosine from the GloVe vector set had a moderately negative relationship with RoBERTa's logits (*r*(2842) = -

0.284, $p < 0.001$), but though the same relationship with human cloze probabilities was significant, it was also small, $r(2842) = -0.090$, $p < 0.0001$. Although the GloVe vectors are well-optimized tools for text representation, they are still unable to describe the knowledge humans use for sentence prediction.

Figure 7

Scatterplots Comparing Performance on the Cloze Procedure



Note: Performance on the cloze procedure by the full RoBERTa model (*Logits*, upper plots) and humans (*Human*, lower plots), plotted against performance on the approximated cloze procedure using RoBERTa's embeddings (*RoBERTa Cosine*, left-hand plots) and the GloVe embeddings (*Glove Cosine*, right-hand plots). Similarity does not have a strong relationship with human cloze probability. The similarity measure from the GloVe embeddings (upper right) had a moderately negative relationship with the model's measure of predictability (*r*(2842) = -.284), which was not replicated with RoBERTa's embeddings (upper left).

Similarity Analyses The poor performance of the static word embeddings cannot simply be due to poor training of the embeddings. Performance of RoBERTa's static embeddings on the WordSim dataset was moderate, Spearman ρ = 0.664. It was lower than GloVe's performance (*ρ* = 0.755) but within the range of human agreement. The lowest Spearman ρ between a single participant and the mean was 0.585, while the highest was 0.837, *M* = 0.758, *SD* = 0.071. The GloVe and RoBERTa embeddings were significantly different, *t*(352) = 4.48, *p* < 0.0001, Cohen's *d* = 0.355, with a fairly small effect size. Although RoBERTa's embeddings are not fully optimized for independent use, they are still within two standard deviations of the mean of human agreement, indicating that they may still be a reasonable test of whether this model's static representation is sufficient for complex language tasks.

Discussion

Leveraging the information captured by static word embeddings against the task of sentence prediction did not achieve the sensitivity to context evident in the full model's performance. This is not because RoBERTa's embeddings were 'stunted' as a result of not receiving adequate feedback during training; when evaluated against human similarity judgments on WordSim353, they were within the range of human agreement. Instead, it must be the case that the knowledge seated in the generative process itself is what drives this sensitivity. Like commonsense knowledge, this is knowledge that only exists in its application.

The slightly negative relationship between RoBERTa logits and GloVe similarity is curious at first glance. Why would semantic relatedness between a word and its context make it more difficult to predict?

The distributional hypothesis of word meaning, which serves as the theoretical foundation of word embedding techniques, may provide insight. Popularized by Firth (1957), it can be summarized as “you shall know a word by the company it keeps,” or that the meaning of a word is revealed through the contexts it regularly occurs in. For instance, although ‘chihuahua’ and ‘labrador’ rarely occur in utterances together, they do often occur in similar contexts, and are therefore recognizable by word embedding algorithms as the same kind of thing. Similarity, therefore, shouldn’t necessarily predict what words go together; in fact, according to the distributional hypothesis, it’s unlikely that highly similar words regularly share the same context.

Given that this relationship was not observed with RoBERTa’s embeddings, the information about this relationship, however indirect, must be also be stored in RoBERTa’s generative architecture.

General Discussion

RoBERTa is a strong model for studying the knowledge underlying linguistic prediction. Not only does its measure of fit describe human agreement, but on the wide variety of linguistic measures we tested, RoBERTa and humans agree about the extent to which they inform predictability. These findings, in conjunction with those from Study 2, suggest that this model is also a potential model for studying commonsense knowledge. Neither the structural nor distributional linguistic measures we looked at addressed the variance in predictability, suggesting that some broader form of knowledge is at play. Further, that human surprisal (interpretable as an inverse measure of fit) and RoBERTa’s logits both showed a graded response across levels of relatedness suggests that this knowledge is well-structured, and that this structure is sensitive to context, both hallmarks of commonsense knowledge. In exploring the structure of this knowledge, we find that static similarity-based representations are unable to explain the context sensitivity exemplified by both RoBERTa and Humans, and conclude that commonsense knowledge cannot be understood in the absence of its application.

The research presented asks many more questions than it answers. Though Study 2 suggested categorical organization, it in fact only directly tested similarity, without testing for clusters of similarity or category boundaries. Is RoBERTa's knowledge actually categorical? If so, this would further strengthen the value of this model as a model of human commonsense knowledge and would provide evidence that categories are necessary for this type of prediction. If not, it would suggest that this type of prediction does not rely on taxonomic categories, but instead alternate semantic structures (i.e., thematic, perhaps). Additionally, word embeddings are known to excel at many tasks of natural language representation. How does the knowledge needed for semantic representation differ from what is used for prediction? Further research should explore both of these points.

What makes a sentence predictable? A lot of research has focused on using biometric responses to understand what cues humans use to predict words, and even more, has used the predictability of a word given a context as a measure in studies of language ability or psychosis, but outside of a general notion of 'constraint,' it's not clear what features actually determine whether a given sentence will be easily predicted. While we capture most non-linguistic knowledge under the umbrella of 'commonsense world knowledge,' what does this knowledge actually consist of? Understanding why humans agree when they do will help us unpack what kinds of knowledge are truly 'common.'

What might such a system of knowledge look like, outside of a computational model? There is growing consensus that episodic and autobiographical memory is constructed online, and with the ever-expanding amount of work under the predictive processing framework, there is some early suggestion that semantic memory may be, too (Fivush & Haden, 2003; Hassabis & Maguire, 2007; Clark, 2015). Not only is there empirical support for predictive neural circuitry across layers of cortex, this evidence has been found in many regions in the brain, leading to the suggestion that prediction is a canonical computation, that the brain is essentially a prediction machine (Hohwy, 2013).

While the early claim that ‘commonsense knowledge is inherently applied’ is intriguing, it is also enigmatic and unfalsifiable. By studying how this model’s knowledge is structured in Study 3, we have gained some insight into how these meta-theories of commonsense knowledge may be understood as more concrete cognitive hypotheses, when placed within a predictive processing framework. Could this predictive circuitry itself store the types of relational knowledge so critical to everyday reasoning? Future research should continue seeking to understand the relative contributions of the static and the dynamic in prediction.

Conclusion

The prevalent encyclopedic approaches to modeling common knowledge seek to explicitly differentiate the ‘items’ of knowledge, all the possible entities, properties, and events that people can know about, from the structures that organize these items in context. After taking a closer look at a system whose intelligence is an emergent property, rather than a set of hand-coded assumptions, we find that this dichotomy is false. Much like autobiographical and episodic memory are not dissociable from the construction process, commonsense knowledge only exists as it is applied.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19-27).

Aitken, K. G. (1977). Using Cloze Procedure as an Overall Language Proficiency Test. *TESOL Quarterly*, *11*(1), 59-67.

Apsari, Y. (2016). Cloze passage in improving student's reading comprehension. *ELTIN Journal, Journal of English Language Teaching in Indonesia*, *4*(2), 53-62.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238-247).

Bollacker, K., Tufts, P., Pierce, T., & Cook, R. (2007). A platform for scalable, collaborative, structured information integration. In *Intl. Workshop on Information Integration on the Web (IIWeb'07)* (pp. 22-27).

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th international conference on neural information processing systems (NIPS '16)* (pp. 4349-4357).

Chen, D. & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014* (pp. 740-750).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Cambria, E., Olsher, D., & Rajagopal, D. (2014, June). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence* (pp. 1515-1521).

Clark, A. (2015). Radical predictive processing*. The Southern Journal of Philosophy, 53*, 3-27.

Carnie, A. (2012). *Syntax: A generative introduction* (Vol. 18). John Wiley & Sons.

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Available online at https://www.english-corpora.org/coca/.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhingra, B., Liu, H., Salakhutdinov, R., & Cohen, W. W. (2017). A comparative study of word embeddings for reading comprehension. *arXiv preprint arXiv:1703.00993*.

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological science, 18*(3), 254-260.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469-495.

Firth, J. R. (1957). 1968. A synopsis of linguistic theory 1930-1955. In F. R. Palmer (Ed.), *Selected papers of JR Firth 1952–1959* (pp. 168-205).

Fivush, R., & Haden, C. A. (Eds.). (2003). *Autobiographical memory and the construction of a narrative self: Developmental and cultural perspectives*. Psychology Press

Gupta, V., Kumar, A., & Bhardwaj, A. (2018). Newsgroup Classification Using CNN and GloVe Embeddings. *International Journal of Applied Research on Information Technology and Computing*, *9*(2), 135-146.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences, 11*(7), 299-306.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Köster, M., Kayhan, E., Langeloh, M., & Hoehl, S. (2020). Making sense of the world: infant learning from a predictive processing perspective. *Perspectives on psychological science*, https://doi.org/10.1177/1745691619895071

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606-608.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211-240.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: toward programs with common sense. *Communications of the ACM*, *33*(8), 30-49.

Lev, G., Klein, B., & Wolf, L. (2015). In defense of word embedding for generic text representation. In *International Conference on Applications of Natural Language to Information Systems* (pp. 35-50). Springer, Cham.

Levy, O., & Goldberg, Y. (2014, June). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302-308).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, & David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60).

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moscovici, S., & Hewstone, M. (1983). Social representations and social explanations: From the “naive” to the “amateur” scientist. In Hewstone, M. (Ed.), *Attribution theory: Social and functional extensions* (pp. 98-125). Blackwell Publishing Limited.

Peelle, J. E., Miller, R. L., Rogers, C. S., Spehar, B., Sommers, M. S., & Van Engen, K. J. (2020). Completion norms for 3085 English sentence contexts. *Behavior Research Methods*, 1-5.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543)

Rezaii, N., Price, B., & Wolff, P. (2020). Natural language processing models in medicine: A paradigm shift. [Manuscript under review]

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Sari, W. A. (2019). *Using cloze procedure technique to increase the students' reading comprehension among the tenth graders at Senior High School 6* (Doctoral dissertation, IAIN Metro).

Schneyer, J. W. (1965). Use of the cloze procedure for improving reading comprehension. *The Reading Teacher*, *19*(3), 174-179.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002, October). Open Mind Common Sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 1223-1237). Springer, Berlin, Heidelberg.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Sternberg, R. J., & Caruso, D. R. (1985). Practical modes of knowing. In Eisner, E.W. (Ed.), *Learning and teaching the ways of knowing, Eighty-fourth yearbook of the NSSE, Part II* (pp. 133-158). University of Chicago

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American psychologist*, *50*(11), 912.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415-433.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development, 3*(1), 1-42.

Toutanova, K., Klein, D., Manning, C., Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003* (pp. 252-259).

Wagner, R. K., & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. *Practical intelligence: Nature and origins of competence in the everyday world*, 51-83.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*(6), 1236–1247. https://doi.org/10.1037/0022-3514.52.6.1236

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*(1), 49-63.

Zettersten, M. (2019). Learning by predicting: How predictive processing informs language development. In Busse, B., Moehlig-Falke, R. (Eds.), *Patterns in language and linguistics: New perspectives on a ubiquitous concept*, *104* (pp. 255-288). De Gruyter.