Getting Situated: Comparative Analysis of Language Models With Experimental Categorization Tasks

Andy Edinger (aedinge@iu.edu)
Program in Cognitive Science & Luddy School of Informatics, Computing and Engineering, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Program in Cognitive Science & Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

Common critiques of natural language processing (NLP) methods cite their lack of multimodal sensory information, claiming an inability to learn situated, action-oriented relations through language alone. Barsalou's (1983) theory of ad hoc categories, which are formed from to achieve goals in real-world scenarios, correspond theoretically to those types of relations with which language models ought to have great difficulty. Recent NLP models have developed dynamic approaches to word representations, where the same word can have different encodings depending on the context in which it appears. Testing these models using categorization tasks with human response data demonstrates that situated properties may be partially captured through semantic analysis. We discuss possible ways in which different notions of situatedness may be distinguished for future development and testing of NLP models.

Keywords: Artificial Intelligence; Language; Situated Cognition; Word Meaning; Distributional Semantics; Categorization;

Introduction

Neural networks operate by learning statistical regularities in data, adjusting weights within the network to minimize output error. An important component of this learning process is the generation of internal hidden vectors corresponding to latent components underlying the inputs (Rumelhart, Hinton, & Williams, 1986). Distributional Semantic Models (DSM's) leverage properties of these hidden vectors for applications in natural language processing, representing semantic information as high-dimensional vectors in which distance between feature vectors is used as a measure of semantic similarity. This approach has proven effective for many natural language processing tasks, with many state-of-the-art models over the past decades employing variations on this approach (Devlin, Chang, Lee, & Toutanova, 2019; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014; Radford, Narasimhan, Salimans, & Sutskever, 2018; Vaswani et al., 2017)

The extent to which latent semantic spaces are able to capture functions of human cognition has attracted previous investigation (Glenberg & Robertson, 2000). However, recent advances in language processing merit revisiting the representational capabilities of such spaces. We explore this question by analyzing performance of several DSM's with human experimental data from cognitive categorization tasks. Grounding assessment of semantic and natural language processing models in data collected from human categorization judgments may yield insight into strategies for improving performance of machine learning models, while simultaneously testing theories of concepts and word meaning in human cognition (Lake & Murphy, 2021).

Table 1: Examples of categories and responses.

Taxonomic Categories:	
Category	Responses
A precious stone	Diamond
	Ruby
	Emerald
	Sapphire
	Pearl
A type of reading material	Magazine
	Book
	Newspaper
	Novel
	Journal
Ad Hoc Categories:	
Category	Responses
Something a clown carries	Balloons
	Pie
	Flowers
	Ball
	Horn
Something children often lose	Teeth
	Toys
	Money
	Homework
	Gloves

Categorization tasks are a diagnostic source of evidence for assessing the representation of semantic information in humans (Barsalou, 1983; Battig & Montague, 1969; Harnad, 2017; Mervis & Rosch, 1981; Van Overschelde, Rawson, & Dunlosky, 2004). Barsalou dissects the notion of natural conceptual categories, delineating common, or taxonomic, categories and ad hoc categories. Common categories comprise typically defined notions of semantic categories, such as "birds" or "fruit." Ad hoc categories, on the

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Table 2: Examples of ad hoc and extended ad hoc category prompts as formatted for BERT's masked language prediction feature. The model replaces the "<MASK>" token with the predicted word of best fit for the sentence.

Original	Extended
<MASK> is something a clown might be carrying.	When the circus came to town and its tent was pitched, the opening act featured an entertaining clown carrying (a) <MASK>.
<MASK> is something that a child often loses.	The child's mother complained that the child was always losing their <MASK> which led to a lot of aggravation for the whole family.
<MASK> is something college students get stolen from them on campus	The freshman student was eager to start college. In their first week at the university they were upset because their <MASK> was stolen.

other hand, consist of "highly specialized and unusual sets of items… created spontaneously for use in specialized contexts" (Barsalou, 1983). Canonical examples of this class include categories such as "things to take on a camping trip" or "things to take from one's home during a fire." Ad hoc categories are typically considered to capture information that is dependent upon world knowledge, obtained from situated experience with objects in a variety of contexts.

There has been considerable interest in ad hoc categories and their apparent dependency upon the ability to understand and simulate real-world situations. For example, in order to realize that "stool" is an example of the ad hoc category "something that can could be used to help you reach a hard-to-reach light bulb", one would need to understand many real-world facts such as the heights of ceilings, the height of people, and the ability of stools to support a standing person. In many cases, these facts need to be generated on the fly because the pre-computed information associated with an object does not suffice to determine whether it matches the goal of an ad hoc category. For example, if one finds oneself on board a sinking cruise ship, then the aptness of a basketball for the ad hoc category "things that can prevent a person from drowning" must be computed in the moment, based on its buoyancy, a property that one may never have contemplated before. Given how important real-world simulations are for people in using ad hoc categories, our primary research question is whether and how can DSM's employ ad hoc categories even though they lack the ability to simulate real-world situations, and are not embedded in a rich, multi-sensory world at all.

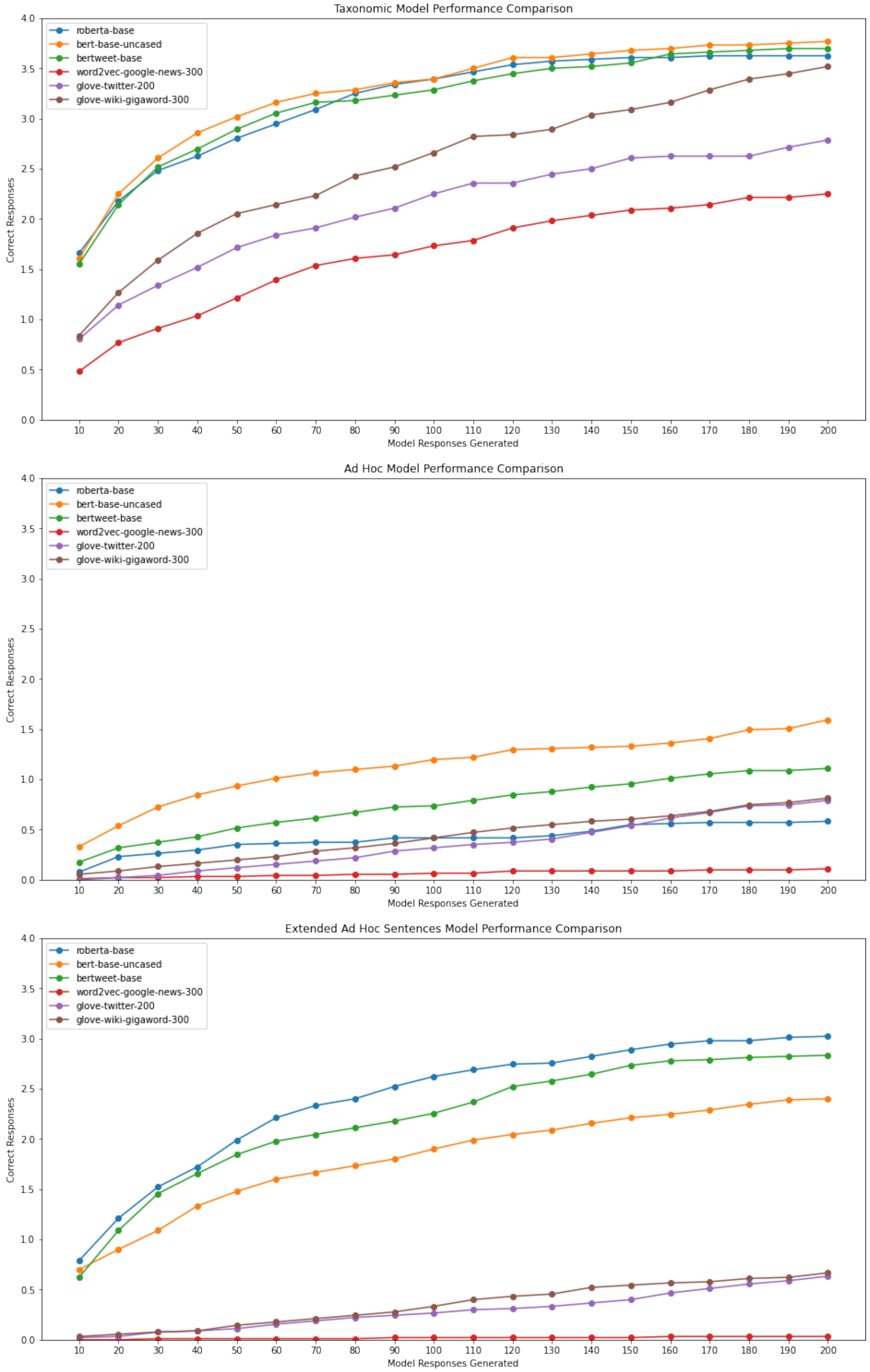
With respect to distributed semantic models, spatial similarity along dimensions of semantic meaning is typically designed to capture associations in language that correspond closely to taxonomic categories. Words that have similar meanings will be closely co-located in feature space - e.g. "dog" and "cat" will fall more closely together than "dog" and "apple". However, ad hoc categories comprise associations based upon situated properties of objects or concepts and the real-world scenarios in which they occur rather than situationally invariant word meanings. Recent work has argued that the gap between NLP models and language use by humans may be attributed to models' lack of integration with rich, multimodal information sources (Birhane, 2021; Dubova, 2022; Lake & Murphy, 2021; McClelland, Hill, Rudolph, Baldridge, & Schütze, 2020). However, there remain fundamental questions about how language models actually use and represent information - questions that the addition of new information sources will do little to answer. By testing model performance relative to human data using categorization tasks that vary in their involvement with situated information, we gain insight into the structure of models' internal representations and the extent to which they may simulate cognitive capabilities that have traditionally been restricted to the domain of situated and embodied cognizing agents.

Methods

To comparatively assess model performance across categorization tasks, we tested six different pre-trained embedding models utilizing three different model architectures on three different categorization tasks.

Datasets

The first dataset we tested was the category norms dataset, a replication of the canonical Battig & Montague norms (Van Overschelde et al., 2004; Battig & Montague, 1969). The original norms were updated for language changes that have occurred due to changes in language usage as a result of phenomena such as cultural trends or commercial products. This dataset comprises the responses of 300 participants to 70 different categories, of which 56 were selected as ideal examples of common categories for the purposes of this study. Identifying data for testing ad hoc categorization presented a somewhat greater challenge. Many of the canonical examples require extended responses for which our models are not well-suited, creating issues for performance evaluation. However, a database of questions and responses from the game show "Family Feud" provided a solution. Questions in Family Feud generally take the form of prompts relating to hypothetical scenarios, with diverse sets of responses


Taxonomic Model Performance Comparison

Ad Hoc Model Performance Comparison

Extended Ad Hoc Sentences Model Performance Comparison

Figure 1: Model response curves for taxonomic (top), ad hoc (middle), and extended ad hoc (bottom) categorization tasks.

crowdsourced from the game show audience. The dataset was reviewed, selecting usable items according to the following criteria to produce 91 items for model evaluation, with examples of items that were determined to fit our criteria provided in table 1:

(1) Responses should be connected through situational or functional properties.

(2) Responses to prompts should be as dissimilar as possible, avoiding questions where a majority of the responses fall into a single category such as "food", "household items", or "occupations".

(3) Items that were not based on situational properties were removed, such as words that begin with "P", words that go with "wax".

(4) Items that had predominantly multi-word or phrasal responses were removed. The model implementations used for testing were not well-suited for multi-word responses, and so these items would introduce undesirable bias and complexity in the task.

In order to test the ability of models to use added contextual information in response generation, a set of extended ad hoc prompts was developed using the items from the family feud dataset. Longer text samples were composed to fit the same prompts and responses, testing the ability of models to use additional descriptive information pertaining to a scenario so as to reduce model ambiguity and produce more accurate results, as observed in table 2.

Models

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a state of the art model for language processing based upon the transformers architecture (Devlin et al., 2019; Vaswani et al., 2017). BERT utilizes an attention mechanism that takes into account not only word co-occurrences but also the position of a word relative to co-occurring words. This means that within a single model, the same word may yield very different encodings depending on the context in which the word appears. This mechanism corresponds theoretically to a key attribute of ad hoc categories in that an object's representation should be dynamically dependent on the context in which it occurs. This stands in contrast to the relatively context-independent associations of taxonomic categories (Barsalou, 1983, 1982). We adapt BERT's masked language prediction function, which takes sample text with a single word removed, or "masked," and yields predictions for tokens that best fit that position in the text. Three different models using the BERT architecture were tested: The BERT base model was pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books, and English Wikipedia (Devlin et al., 2019; Zhu et al., 2015). BERTweet is a model designed to capture linguistic patterns in social media usage, was trained on 850m English-language tweets(Nguyen, Vu, & Tuan Nguyen, 2020). RoBERTa is a replication of BERT with reoptimized pretraining and hyperparameter tuning processes(Liu et al., 2019).

Word2Vec is a neural network model that uses word context to learn vector representations of words (Mikolov, Sutskever, et al., 2013). Two different learning algorithms may be used. In the Continuous Bag of Words (CBOW) algorithm, the model attempts takes a word and attempts to predict its neighbors in context. In the Skip-Gram algorithm, context is used to predict the current word. In both cases, weights are adjusted such that pairs of words that commonly occur with many of the same words will tend to be spatially close within the final learned vectors. Importantly, in contrast to BERT, after training occurs each word will be associated with only a single embedding vector. One Word2Vec model was tested: The Word2Vec "google-news-300" model, pretrained on the Google News dataset and implemented via Gensim (Rehurek & Sojka, 2011).

GloVe, which stands for Global Vectors, is a statistical model that uses word co-occurrences counts in the entire training corpus combined with dimensionality reduction techniques to produce word vectors with features corresponding to statistical regularities in the co-occurrence counts. Similar to Word2Vec, GloVe produces static word vectors after the training process. However, GloVe thus places greater emphasis on global word contexts than Word2Vec's local context analysis. Two GloVe models were tested as implemented via Gensim: "glove-twitter-200", trained on 2 billion tweets, and "glove-wiki-gigaword-300", trained on English wikipedia and news articles (Pennington et al., 2014; Rehurek & Sojka, 2011).

These models were selected so as to include a cross-section of widely-used state of the art language models, varying across models and pretraining data to mitigate influences of idiosyncrasies in particular training sources. Some prominent models were excluded, such as the GPT series (Radford et al., 2018), as their word-generating structure was not amenable to completing the prompts in such a way that one-to-one comparisons with other models would be meaningful.

Results

For each item in the ad hoc dataset, seven responses were included. The common category norms listed all responses that participants provided, ordered by the proportion of participants that gave a response. In order to directly compare model performance across tasks, we used only the top seven participant responses for each category in testing our models. Each model was queried to produce up to 200 best-fitting items for each category in steps of 10, and the number of correct responses was recorded as a function of the number of responses given. The response curves are plotted as shown in Figure 1.

We observe better performance in the BERT models than the GloVe and Word2Vec models for the three tasks across nearly all quantities of responses given, with the exception of the Roberta model for the ad hoc task, which was surpassed by the GloVe models when greater than 100 responses were tested. For any number of responses given, the best BERT model outperforms the best of the GloVe and Word2Vec mod-

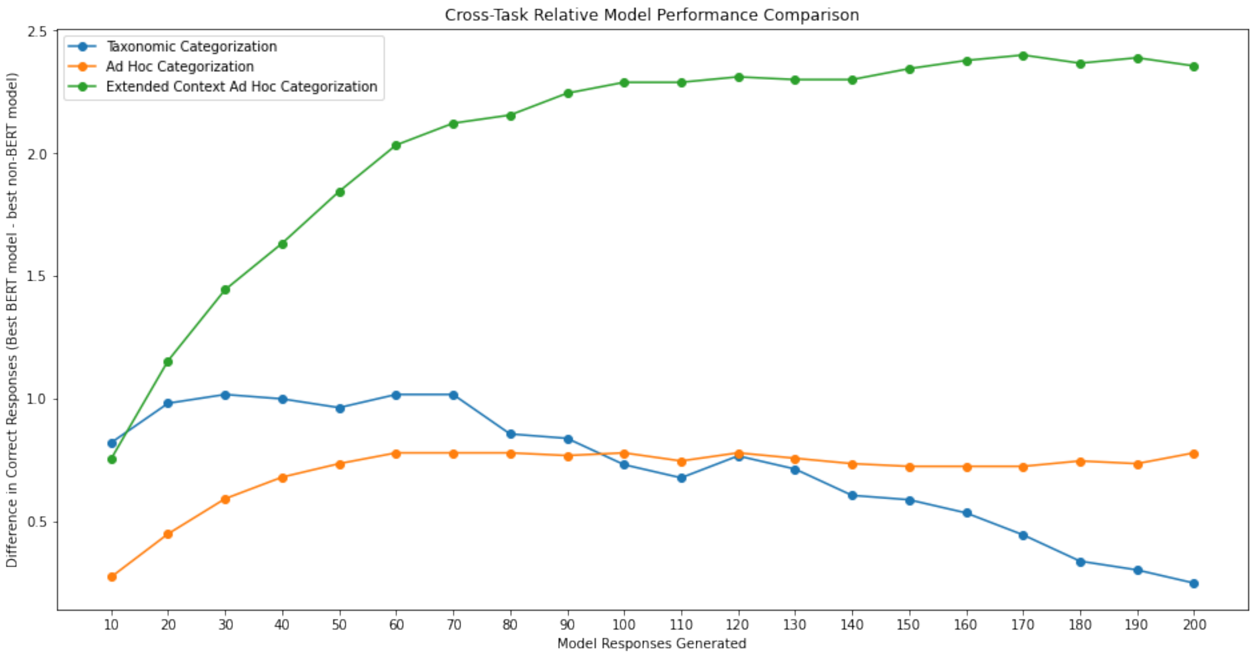


Figure 2: Relative model performance for each task, computed as the best BERT model score minus the best of the GloVe and Word2Vec model scores.

els, as shown in Figure 2 The shape of the response curves yields further insight into the strength of the BERT models, with sharper growth at lower numbers of responses. This indicates that the BERT models are able to capture a more coherent, clustered category representation than the alternatives.

There was greater difference in model performance for the ad hoc task than for the taxonomic categorization tasks, and the most pronounced difference was observed with the extended task. The GloVe and Word2Vec models showed nearly zero performance improvement with the added context, while the BERT models exhibited significantly improved performance. This supports the hypothesis that BERT's attention mechanism is able to leverage contextual information in the identification of properties of situated concepts.

It should be noted that absolute performance of all models was quite low. At best, the models produce an average of approximately 2 correct responses out of 20 generated for the taxonomic task and only about 1 out of 20 for the ad hoc tasks, with proportional accuracy decreasing relative to the total number of responses given. However, these models are not trained or fine-tuned in any way for this task, only trained to capture general semantic information on broad datasets. GloVe and Word2Vec average less than 0.5 correct responses out of 200 for the ad hoc tasks, so BERT's modest averages of 1.5 and 3 for the ad hoc and extended ad hoc tasks, respectively, constitute a significant performance increase over baseline. Furthermore, benefits of contextualization for the BERT models are likely underestimated due to the effect of extended contextualized prompts restricting the number of responses that fit the specific context. For example, the prompt ”¡MASK¿ is something that a child often loses” might be equally well fit by ”teeth” or ”homework” while the extended prompt (table 2) would bias towards ”homework”.

Discussion

Many common critiques of natural language processing models revolve around their disembodied approach to evaluating meaning, arguing that there are facets of conceptual meaning in human cognition that cannot be captured by language inputs alone. The role of situational understanding as a cornerstone of cognition is not a new concept, and there is strong evidence to suggest that people's understanding of written language is grounded in situated knowledge. Modeling situated knowledge will be integral for developing adaptable language models that closely resemble human language processing. McClelland et al. argue that further progress in language modeling will necessitate the treatment of language as part of a larger communication system, and Bisk et al. argue that an understanding of action-oriented categories requires a cognitive agent to be able to participate in situated action with its environment to be able to discover such categorical associations. Proposed solutions to these challenges often revolve around the integration with simulations of physical characteristics or other sources of information to produce multimodal conceptual representations. While such improvements to machine learning systems show potential for expanding the representational capabilities of NLP models, these approaches alone will not bridge the gap between current NLP capabilities and the conceptual flexibility of human cognition.

Of course, concepts in cognition as evoked through language and communication are irrevocably tied to properties of those concepts across the variety of modalities in which they exist. The use of the word "dog" may produce general associations of any sensory or conceptual understanding in which humans participate, in addition to specific associations grounded in real-world memories or experiences. However, an understanding of the concept of a "dog", while tied to all of these, might also be arrived at through purely communicative means without real-world physical experiences with a dog. This is a common occurrence, as descriptions of hypothetical or impossible scenarios, or metaphysical concepts may be relayed through language to arrive at a shared understanding of the term. A primary function of language is the transfer of shared symbols to communicate the properties of a given scenario, a process in which humans engage daily, inferring deep relational properties from semantic information. This process is demonstrated, albeit roughly, by the models tested in this study. Through the analysis of semantic information they are able to partially capture the association of disparate concepts through situated, action-oriented properties.

The distinction between taxonomic and ad hoc categories corresponds strongly with the strengths and weaknesses commonly associated with NLP models, and testing models on these types of tasks creates opportunities for critically analyzing the representative capabilities of language models, allowing for the development of actionable strategies for improving the ways in which information is processed. The results of this study show that models are continuing to improve in their ability to infer situated properties from semantic context, indicating that there is yet ground to be gained from refining the way in which these implicit properties may be inferred. Identification and construction of further tasks whose successful completion relies on the utilization of situated relations and action-oriented category information, and that allow for variation in those properties, create opportunities to test model capabilities and identify the specific shortcomings in these models. These strategies will allow us to progress past the broad-strokes critiques of language models, allowing us to distinguish between different senses of situated language.

None of the models tested here incorporate multimodal inputs, relying only on language from written text as input, and are far from anything that might be considered an agent-environment model. In this sense, none of the models are perceptually situated. Even so, there were significant differences in model performance, particularly on the ad hoc categorization tasks. One of the primary distinctions between the NLP approaches tested here is whether a model assigns a single vector representation for an item, as in GloVe and Word2Vec, or develops a situationally specific representation for a word based on other words in its immediate textual context, as in BERT. In this sense, BERT is able to capture situated properties as context-dependent encodings. Furthermore, developing context-dependent encodings yields particular advantages for ad hoc category prompts that were designed to be more situationally specific and, crucially, offers greater benefits for ad hoc than taxonomic categories. Despite underwhelming absolute performance, the relative performance gain in capturing situated information demonstrated by BERT's encoding scheme indicates the need for further development of language models - not just in the kinds of information that are provided, but in the ways in which information is gathered and represented.

Acknowledgments

This work is supported by fellowship funding to Andy Edinger from National Science Foundation Research Traineeship Grant 1735095 "Interdisciplinary Training in Complex Networks and Systems."

References

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, 2–93.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes*, *18*(5-6), 513–562.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, *80*(3, Pt.2), 1. (Publisher: US: American Psychological Association)

Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, *27*(1), 44–61.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... Turian, J. (2020). Experience Grounds Language. *arXiv:2004.10151 [cs]*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Dubova, M. (2022). Building human-like communicative intelligence: A grounded perspective. *Cognitive Systems Research*, *72*, 63–79.

Family Feud Question Database - Google Drive. (n.d.). Retrieved 2022-02-02, from https://docs.google.com/spreadsheets/d/ 1y5TtM4rXHfv9*BktCiJEW*621939*RzJucXxhJidJZbfQ /htmlview*

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Harnad, S. (2017). To cognize is to categorize. In *Handbook of Categorization in Cognitive Science* (pp. 21–54). Elsevier.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. *arXiv:1904.12584 [cs]*.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(1), 89–115.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv:1310.4546 [cs, stat]*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2016). Linguistic regularities in continuous space word representations. , 6.

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Online: Association for Computational Linguistics.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. , 12.

Rehurek, R., & Sojka, P. (2011). Gensim-python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088).

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., … Bengio, Y. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv:1502.03044 [cs]*.

Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. *arXiv:1905.04562 [cs]*.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv:1506.06724 [cs]*.