Semantic features of object concepts generated with GPT-3

Hannes Hansen (hansen@cbs.mpg.de)
Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a 04103 Leipzig Germany

Martin N. Hebart (hebart@cbs.mpg.de)
Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstraße 1a 04103 Leipzig Germany

Abstract

Semantic features have been playing a central role in investigating the nature of our conceptual representations. Yet the enormous time and effort required to empirically sample and norm features from human raters has restricted their use to a limited set of manually curated concepts. Given recent promising developments with transformer-based language models, here we asked whether it was possible to use such models to automatically generate meaningful lists of properties for arbitrary object concepts and whether these models would produce features similar to those found in humans. To this end, we probed a GPT-3 model to generate semantic features for 1,854 objects and compared automatically-generated features to existing human feature norms. GPT-3 generated many more features than humans, yet showed a similar distribution in the types of generated features. Generated feature norms rivaled human norms in predicting similarity, relatedness, and category membership, while variance partitioning demonstrated that these predictions were driven by similar variance in humans and GPT-3. Together, these results highlight the potential of large language models to capture important facets of human knowledge and yield a new approach for automatically generating interpretable feature sets, thus drastically expanding the potential use of semantic features in psychological and linguistic studies.

Keywords: semantic features; conceptual knowledge; natural language processing; GPT-3

Introduction

A central aim in the cognitive sciences is to understand the nature of human conceptual knowledge. This knowledge is often characterized through semantic features, which form a set of minimal semantic descriptions of concepts and which have been at the heart of much theorizing about categorization (Nosofsky, 1986; Rosch, 1973), semantic memory (Murphy, 2004), and semantic processing more generally (Cree & McRae, 2003). For example, the concept *car* can, among others, be described by the features *is a vehicle* and *has four wheels*. The relationship of this concept to other concepts can then be quantified by evaluating the similarities and differences to the features of other concepts.

A common approach for attaining semantic features of concepts is to instruct humans to list properties for a given concept, for example by asking *what are the properties of a cow?*. The popularity of such empirically-generated semantic features has led researchers to create semantic feature production norms for a larger number of concepts (Devereux, Tyler, Geertzen, & Randall, 2014; McRae, Cree, Seidenberg, & McNorgan, 2005), which have been invaluable for improving our understanding of semantic representations. At the same time, the impact of these norms has remained constrained to the set of concepts and features that have been made publicly available. Creating new norms requires collecting responses from hundreds of participants and necessitates extensive manual curation by researchers, inherently restricting the scope of such approaches. If there was a computational model that contained the knowledge sufficient for generating feature norms of similar quality to humans, this would drastically expand the possible scope of research with semantic features in the study of conceptual knowledge.

In recent years, there have been strong advances in the field of natural language processing. So-called transformer models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) or GPT-3 (Brown et al., 2020), often approach human-level performance in diverse language understanding tasks (Floridi & Chiriatti, 2020; Wang et al., 2018) and can even be used to produce prose that can be difficult to distinguish from human-generated text (Dale, 2021). While the general linguistic understanding and reasoning ability of these models are still far from perfect (Marcus, 2020), they may offer a valuable computational tool for automatically producing semantic features for an arbitrary number of concepts (Derby, Miller, & Devereux, 2019), thus leveraging the statistical structure of knowledge present in their immense training text corpora for addressing central questions in semantic cognition research (Bhatia & Richie, 2022).

Here we tested the degree to which recent transformer models can be used for automatic production of semantic features and whether the produced features mirror those found in humans. To this end, we used GPT-3 to generate a semantic feature norm for 1,854 diverse concepts of concrete objects. We chose concrete objects for two reasons. First, concrete objects have been used in much research on conceptual representations and are indeed used in several existing feature production norms, thus providing a valuable human baseline to relate our results to. Second, similarity ratings for these 1,854 objects have recently become available (Hebart, Zheng, Pereira, & Baker, 2020), providing a broad test case for validating these features with existing similarity ratings.

Related research

Several previous studies have investigated the use of corpus-based language models to produce sets of features mirroring

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

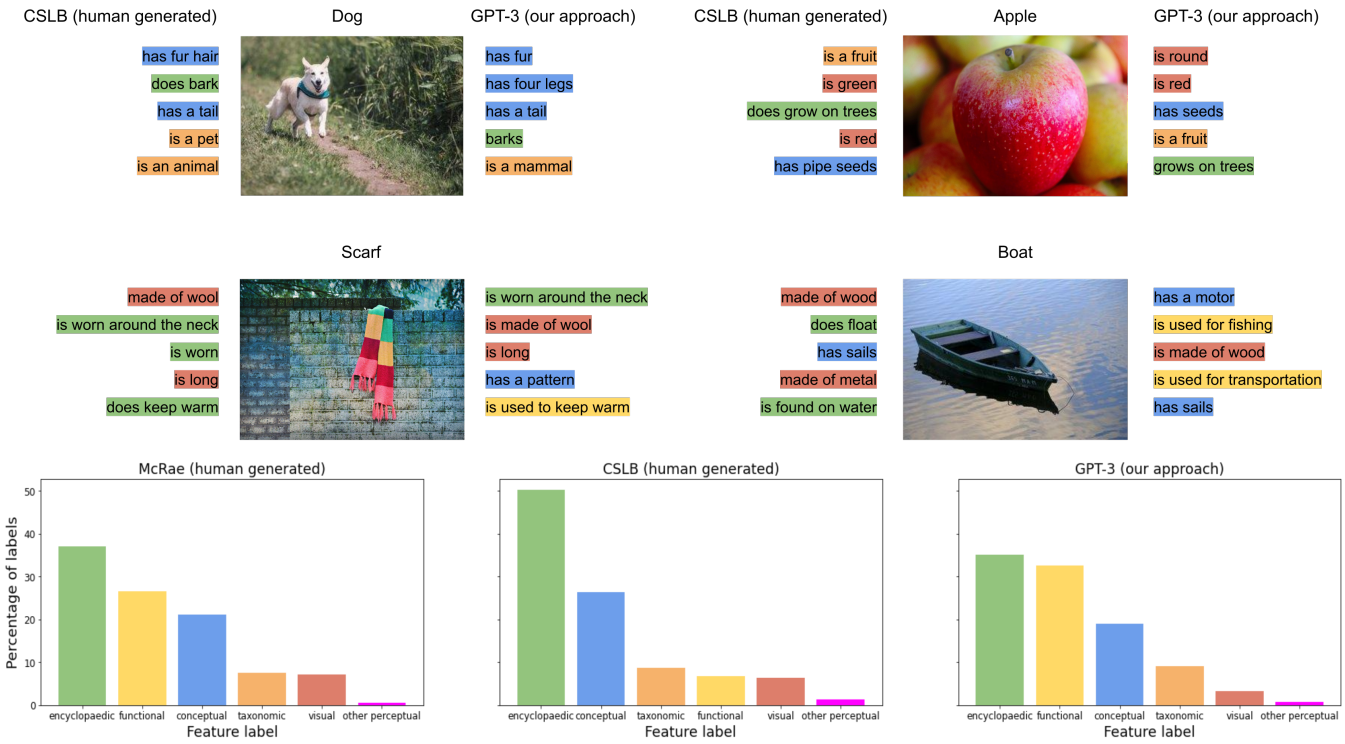

Figure 1: Comparison of features from the CSLB norms and features generated using GPT-3, including feature examples (top) and feature label distributions (bottom). Please note that images were not part of the study and are used for illustration only. Bootstrapped confidence intervals for feature label distributions were too small to be displayed.

Table 1: Descriptive statistics of human and GPT-3 generated feature norms

Feature	CSLB	McRae	GPT-3 (preprocessed)	GPT-3 (preprocessed + filtered)
Number of concepts	638	541	1,854	1,854
Total number of features	22,667	7,259	189,126	124,569
Number of unique features	5,929	2,524	63,467	11,683
Number of features per concept	35.52	13.42	102.06	67.35
Share of unique features to all features	26.16%	34.77%	33.55%	9.37%

human conceptual knowledge. Static word embeddings, including *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013) or *GloVe* (Pennington, Socher, & Manning, 2014) are trained on lexical co-occurrences in large text corpora and provide decent predictions of human similarity ratings (Hill, Reichart, & Korhonen, 2015). However, the features of these models typically lack interpretability (Subramanian, Pruthi, Jhamtani, Berg-Kirkpatrick, & Hovy, 2018). Rubinstein, Levi, Schwartz, and Rappoport (2015) used word embeddings to directly predict a small set of semantic features, concluding that distributed language models may be better at capturing taxonomic than attributive features. Făgărăşan, Vecchi, and Clark (2015) used partial least squares regression to map word embeddings to feature norms, with a more recent approach using a non-linear mapping based on a multilayer perceptron (Li & Summers-Stay, 2019). Derby et al. (2019) proposed *Feature2Vec*, a method which combines the information from word embeddings with human feature norms by projecting the features into the word embedding space. Despite good overall performance, these methods rely on a fixed feature vocabulary, making it only possible to generate features for new concepts but not completely new features.

More recently, Bhatia and Richie (2022) investigated the use of transformer models to model human conceptual knowledge by finetuning a pretrained BERT model (Devlin et al., 2019) on a broad set of features and probing the model whether the concept-feature pairs were correct. The model correctly predicted concept features even outside of the training set with good accuracy, providing an important step towards general purpose feature generators. However, this method still requires probing existing feature-concept pairs, rather than generating new features. Thus, it remains unknown to what degree it is possible to generate arbitrary features for arbitrary concepts. Finally, Bosselut et al. (2019)

and Petroni et al. (2019) proposed models of commonsense knowledge based on GPT and BERT. Models are finetuned on subject-relation-object triplets, with the task of predicting the object (e.g. *mango-IsA-fruit*). While these models can partially generate features for concepts, they are limited in that they require a prori specification of relations.

Methods

Feature collection from GPT-3

We probed GPT-3 (DaVinci version) to generate semantic features for 1,854 object concepts from the THINGS database (Hebart et al., 2019), using a text completion task. To instruct GPT-3 on this task, we presented it with a question about an object (e.g. *What are the properties of a chair?*), had it generate an answer, and replaced this answer with features from the McRae feature norm (McRae et al., 2005) which served as ground truth (e.g. *It is furniture, it is made of wood, etc.*). When continuing this process, generated features appeared to no longer improve in quality after three question and answer sequences, so we chose this number as a trade-off between monetary cost and performance. Once GPT-3 was primed on the task, this was followed by an open question for each of the 1,854 object concepts, after which the answer was collected and both the answer and the question deleted to keep the context static across all trials. To reduce bias induced by specific concepts and associated features and to better mirror experimental approaches that merge data across human participants, we collected 30 runs, each time using a different set of example concepts. In cases where a concept occurred multiple times with different meanings (e.g. bat as animal or bat as sports item), we added a superordinate category to the training example in parantheses.

Preprocessing of features

For better comparability to humans, we preprocessed and normed generated features. We automated preprocessing by using part of speech tagging with the Python library *spacy*[1] and applying a set of preprocessing rules (see below).

The produced answers consisted of lists of features which were split at commas in order to attain raw features. Next, raw features that were classified as nonsensical, consisted of a single word, or were tautological (e.g. *a rose is a rose*) were removed. In addition, features not beginning with a pronoun (e.g. *green color* rather than *it has green color*) or features containing non-ASCII characters (*it has* ©) or question marks were removed. Finally, qualifier adverbs (e.g. *usually* or *really*) were removed, in line with previous approaches (Devereux et al., 2014). This affected 0.75% of all features.

Next, long features with a subordinate clause (*which, that, when, if, but*) were shortened by removing the subordinate part. For example, a feature like *it is a car that drives* was shortened to *it is a car*. This affected 1.83% of all features.

These clean and concise features were furthermore normed. Nested features, containing multiple units of information, were extracted. For example, a feature like *it is a big tree* was decomposed automatically into *it is a tree* and *it is big*. A feature like *it is blue and green* was decomposed into *it is blue* and *it is green*. Next, features containing synonyms, e.g. *it is a car* and *it is an automobile* were collapsed using Wordnet synsets in the Python library *nltk*[2] by choosing the more frequent synset. However, to avoid merging non-synonymous words, two words were only considered as synonyms if their most frequent synset was the same. 4.3% of all features were replaced.

Filtering

The generated feature norm partially consisted of overly sparse features, with many features that were unique to individual concepts. However, a smaller set of features is desirable both for reasons of better interpretability and for reduced computational cost. Therefore, we removed features that occurred very infrequently within each concept. This also made the feature norm more comparable in size to human generated norms. To identify a cutoff, we plotted the number of unique features after removal of infrequent features and chose the elbow point at k=4. Importantly, while this step strongly reduced the number of unique features (see Table 1), it did not affect performance in validating the norm.

Results

Comparison with human feature norms

As a first analysis, we compared the GPT-3 feature norms with human feature norms McRae (McRae et al., 2005) and CSLB (Devereux et al., 2014), using descriptive statistics. We did not use the Buchanan norm (Buchanan, Valentine, & Maxwell, 2019) as it was composed only of associative features. As seen in Table 1, the preprocessed GPT-3 feature norm was computed for a larger number of concepts, leading to a larger number of total features and more unique features. Without filtering, a similar share of unique features can be seen as in the McRae norm, indicating a similar level of redundancy of features across concepts as found in humans. After excluding rare features, only around 9% of all features remained unique. Of note, GPT-3 produced a much larger number of features per concept, indicating that human feature production may be limited to a sparser set of features than those found in GPT-3.

Label distribution

Figure 1 (top) shows several example concepts with their five most frequent features. The results indicate many similarities and no obvious errors in the types of labels assigned by GPT-3. To directly compare the distribution of semantic features between humans and GPT-3, samples of features from the CSLB feature norms and GPT-3 were labeled into the categories taxonomic, visual perceptual, other perceptual, conceptual, functional and encyclopedic. We use a slightly different naming scheme of feature types to McRae and CSLB to

[1] https://spacy.io/

[2] https://www.nltk.org/

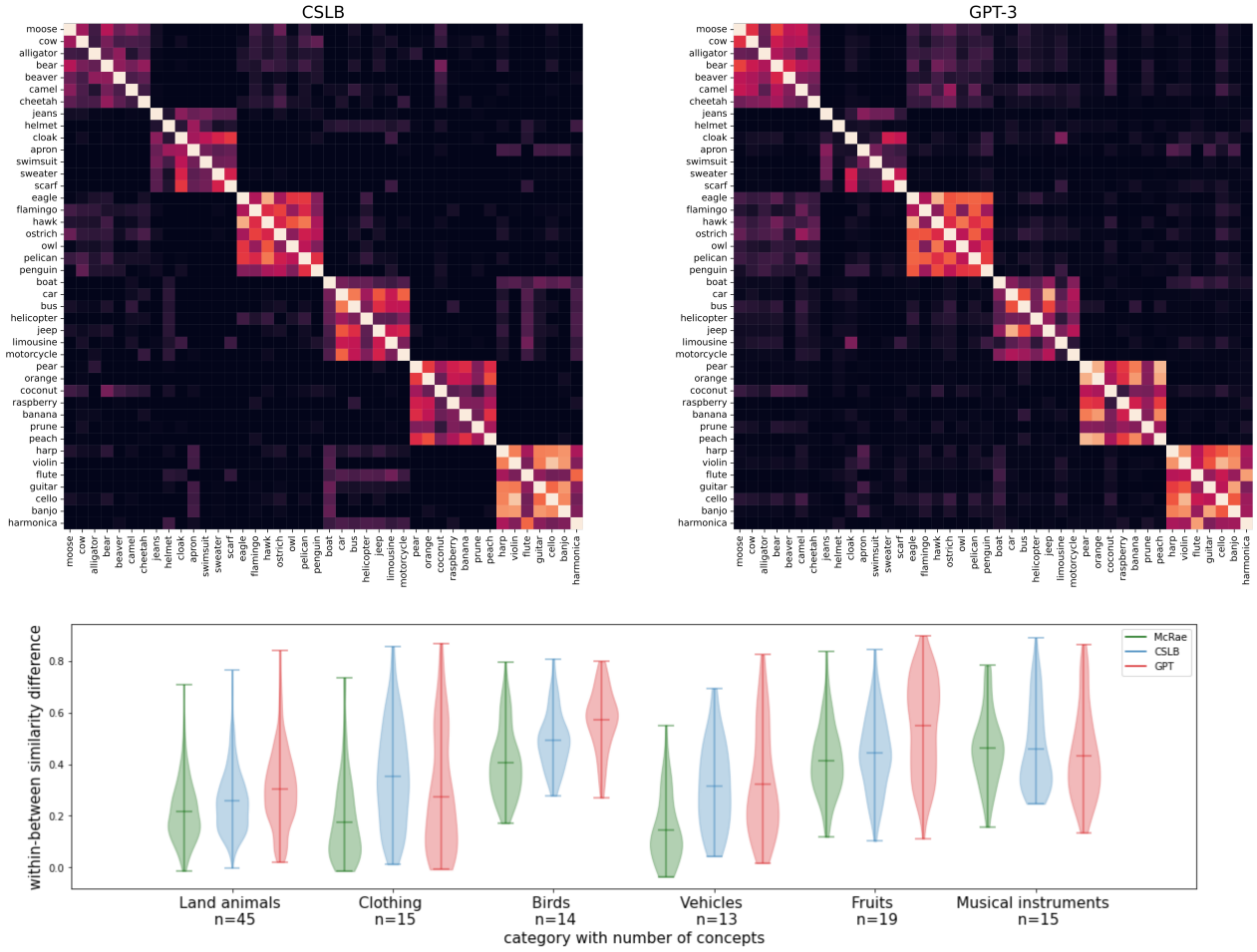

Figure 2: Representational similarity matrices for 6 concepts in the categories land animals, fruits, vehicles, fruits, birds and clothing using the CSLB and GPT-3 feature norms (top) and pairwise cosine similarities per category (bottom).

allow for more fine-grained differences between conceptual, functional, and encyclopedic features. To estimate the distribution of features in all three norms, we randomly sampled 500 features from the 317 concepts shared between CSLB, McRae, and GPT-3 and 500 features from concepts outside of the intersection. The corresponding feature labels were assigned manually.

The results in Figure 1 (bottom) show that humans and GPT-3 mostly rely on encyclopedic features and less on perceptual features. GPT-3 contained a larger number of functional features as compared to the CSLB norm. However, the differences in the distribution to the McRae norm, which GPT-3 had been trained on, were less prominent, indicating that differences of GPT-3 to CSLB may be driven more by differences in populations for the creation of human norms or norming processes, rather than intrinsic differences in representations. Overall, this analysis yielded no obvious differences to human norms in the types of semantic features that were generated.

Table 2: Similarity and relatedness prediction

Pearson correlation	McRae	CSLB	GPT-3
MEN (n=55 word pairs)	0.77	0.77	0.79
Simlex-999 (n=26 word pairs)	0.60	0.63	0.62
THINGS (n=317 concepts)	0.56	0.63	0.62

Prediction of category structure based on feature similarity

Next we tested the degree to which GPT-3 generated feature norms produced reasonable category structure and how they compared to existing norms. For better comparability, several analyses were conducted in a similar fashion to those found for the creation of the CSLB norm (Devereux et al., 2014). For a fair comparison, we restricted our analyses to the 317 concepts shared between the McRae, CSLB and GPT-3 feature norms. Similarities were based on the cosine similarity of the concepts across feature vectors composed of the pro-

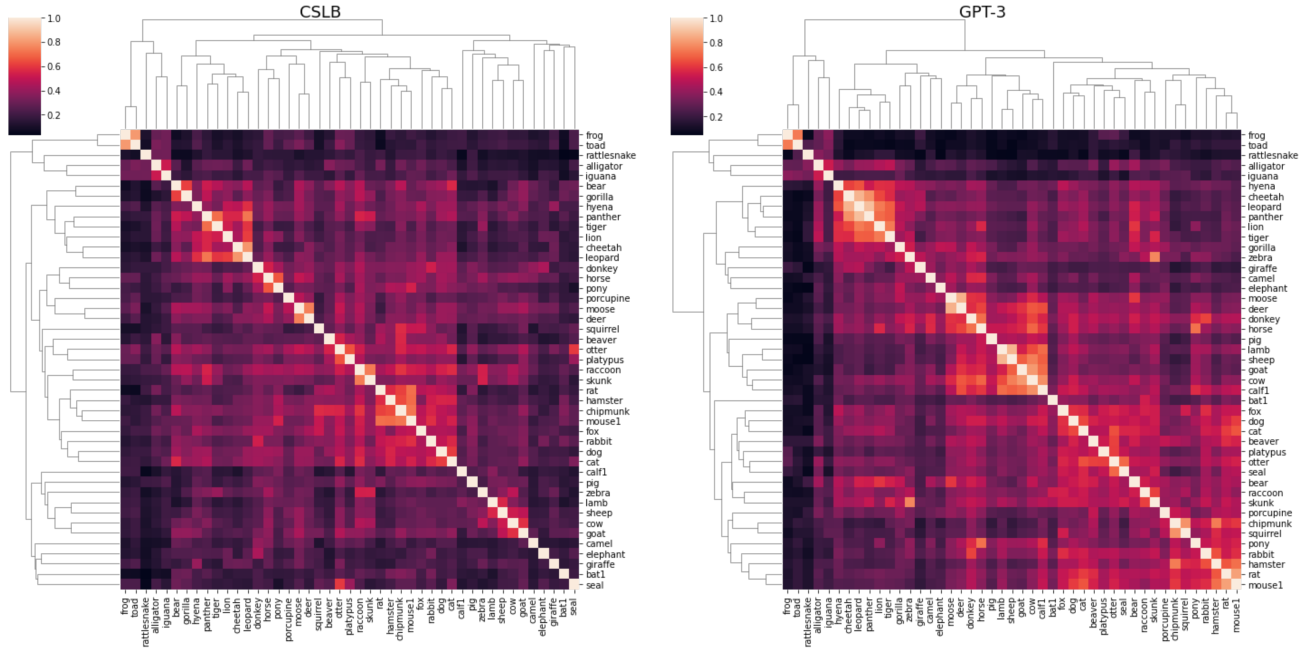

Figure 3: Hierarchical clustering for 45 land-animal concepts using the CSLB and GPT-3 feature norms, highlighting similarities and differences for within category representational structure.

duction frequency per feature, for human norms across participants and for GPT-3 across the 30 runs for which it had been primed with different examples.

First, we inspected the superordinate category structure. Mirroring the results of (Devereux et al., 2014), we based these analyses on six exemplars of the six categories *land animals*, *clothing*, *birds*, *vehicles*, *fruits*, and *musical instruments*. If the produced features prove to be useful for predicting high-level category structure, we would expect consistently high within-category similarity and low between-category similarity. A visualization of the similarity structure of the CSLB norm with the GPT-3 generated norm is shown in Figure 2 (top). As is evident from the figure, both norms show excellent category structure, clearly distinguishing high-level categories from each other. To quantify similarities and differences between norms, we expanded the set of concepts available within each superordinate category to the 317 concepts and computed the mean within-category similarity minus the mean between-category similarity of each concept. The results are shown in Figure 2 (bottom). Overall, GPT-3 showed comparable or clearer category structure than human norms, with the only exception of *clothing*. Together, these results demonstrate that high-level category structure can be extracted from GPT-3 generated feature norms, with similar performance to humans.

Beyond between-category effects, we explored whether the within-category similarity structure was meaningful in GPT-3 generated norms. To this end, we performed hierarchical clustering of the 45 land animal concepts. Overall, we expected high similarities between all animals but also more fine-grained differences. The results comparing CSLB with GPT-3 norms are shown in Figure 3. Overall, the similarities between animals were higher in the GPT-3 norm than CSLB. A clear clustering of highly similar animals is found in both matrices (e.g. *lamb* and *sheep*). However, the subordinate category structure was slightly different between GPT-3 than CSLB. For example, farm animals clustered in GPT-3 but were more distributed in CSLB, while dangerous animals clustered more closely in CSLB. In sum, while the overall category structure was similar, there were fine-grained differences in the representations derived from semantic features through GPT-3 as compared to the CSLB norm.

Prediction of similarity and relatedness ratings

Predictions of similarity and relatedness tasks are often seen as a gold standard for evaluating the relationship between semantic features and our conceptual representations. To identify the degree to which GPT-3 generated norms could be used to predict similarity and relatedness ratings, we used the overlap between McRae, CSLB, and GPT-3 generated norms with two existing datasets commonly used as natural language processing benchmarks: The MEN dataset (Bruni, Tran, & Baroni, 2014) was used for word relatedness and the Simlex-999 dataset (Hill et al., 2015) for word similarity. We intended to include other common benchmark datasets but did not find sufficient overlap in concepts. Beyond these datasets, we used human similarity scores from THINGS (Hebart et al., 2020), which were available for all included objects. Similarities were compared by using the Pearson correlation of the

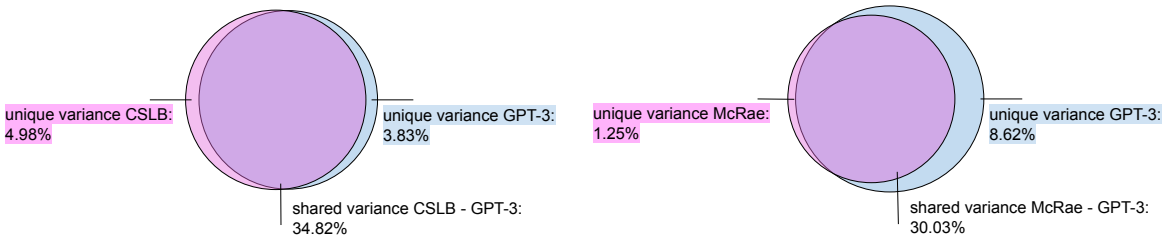

Figure 4: Variance partitioning of human similarity judgements, demonstrating strong overlap between human generated and GPT-3 generated norms in predicting human similarity judgments.

flattened lower triangular part of the similarity matrices.

The overlap in the number of concepts as well as the performance for McRae, CSLB, and GPT-3 generated norms are found in Table 2. While CSLB performed better than McRae for the prediction of THINGS, across all three datasets, the performance of CSLB and GPT-3 was comparable. This indicates that the representations derived from GPT-3 generated norms were of similar quality as those found in human norms. What is left open by these correlations is whether the predictions of similarity were based on similar information in McRae, CSLB, and GPT-3, or whether GPT-3 had access to other information not reported by humans. To address this question, we carried out variance partitioning, identifying the unique and shared variance components explained in the THINGS similarity dataset. The results of this analysis are shown in Figure 4. Overall, there was strong overlap in the explained variance between McRae and GPT-3 as well as CSLB and GPT-3, with GPT-3 subsuming much of the variance explained by McRae, while explaining very similar portions of variance than CSLB. Overall, this result demonstrates that, indeed, GPT-3 is relying on similar information as CSLB for predicting human similarity.

Discussion

Overall, the results demonstrate that GPT-3 can be used for automatically generating semantic feature norms for a large number of concrete objects. Frequently produced features were meaningful and comparable in distribution to those found in humans. Further, the results showed that superordinate categories were well identified by the resulting features and that within-category structure was reasonable. Predictions of similarity and relatedness were comparable to humans and relied on similar information. Overall, this demonstrates that GPT-3 can serve as an effective automatic feature generator, opening up an efficient and rapid approach to generate large numbers of features for diverse concepts.

Of note, there were some differences to humans. Overall, GPT-3 produced a much larger number of features, yet reducing this set through filtering led to very similar results in the prediction of similarity ratings. Further, the fine-grained similarity structure was slightly different to CSLB. Future work is needed to investigate these differences and the degree to which they are related to differences between representations in humans and GPT-3, or whether they were driven by the norming process itself.

GPT-3 was primed using only three examples, which highlights the simplicity of our proposed approach but which may also introduce bias. Rather than reflecting the knowledge available to the model, it may in fact mimic the process of feature generation produced by humans in the three examples it was provided. Other, more effective priming procedures may be discovered in the future that constitute less bias and are better at revealing the knowledge available in such models. Future work may also investigate the influence of model complexity on feature generation.

Nevertheless, the fact that GPT-3 and humans relied on similar information for predicting similarity ratings is relevant in its own respect, indicating that GPT-3 may indeed contain a lot of information useful for modeling important aspects of conceptual knowledge (Bhatia & Richie, 2022). As a consequence, it may be possible to use GPT-3 to generate other types of norms, for example ratings of animacy, graspability, or size (Grand, Blank, Pereira, & Fedorenko, 2022). We did not test whether GPT-3 was able to produce meaningful features for more abstract concepts or verbs, which is an important avenue for future research. While better computational models for producing semantic features may exist, our work demonstrates that it is already possible to create semantic features for concrete concepts with human level performance in predicting similarity ratings using GPT-3.

Conclusions

Here, we introduced a new approach for automatically generating semantic features for diverse concepts using the recent transformer-based model architecture GPT-3. Our results demonstrate that recent large language models are indeed able to accurately reflect important aspects of human conceptual knowledge. The approach opens the door to automatic feature generation for arbitrary concepts, thus widening the potential scope of semantic features for research in psychology and linguistics. To promote the general use of this method and results, the GPT-3 generated raw data as well as the final feature norm of the 1,854 object, including the code to probe GPT-3 and to preprocess and filter raw features, are made publicly available[3].

[3]https://github.com/ViCCo-Group/semantic_features_gpt_3

References

Bhatia, S., & Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological Review*.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019, July). COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4762–4779). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1470 doi: 10.18653/v1/P19-1470

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*, 1–47.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849–1863.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of experimental psychology: general*, *132*(2), 163.

Dale, R. (2021). Gpt-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118.

Derby, S., Miller, P., & Devereux, B. (2019, November). Feature2Vec: Distributional semantic modelling of human property knowledge. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 5853–5859). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1595 doi: 10.18653/v1/D19-1595

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, *46*(4), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423 doi: 10.18653/v1/N19-1423

Floridi, L., & Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694.

Făgărăşan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th international conference on computational semantics* (pp. 52–57).

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 1–13.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Li, D., & Summers-Stay, D. (2019). Mapping distributional semantics to property norms with deep neural networks. *Big Data and Cognitive Computing*, *3*(2), 30.

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

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

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

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

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 2463–2473). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1250 doi: 10.18653/v1/D19-1250

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language* (pp. 111–144). Elsevier.

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 726–730).

Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., & Hovy, E. (2018). Spine: Sparse interpretable neural embeddings. In *Thirty-second aaai conference on artificial intelligence.*

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-5446 doi: 10.18653/v1/W18-5446