Eliciting Complex Relational Knowledge from Masked Language Models

Arun Sundaresan[1], Ming Hsu[2], Zhihao Zhang[2]
[1]UC Berkeley
[2]Haas School of Business, UC Berkeley
arun.sundaresan@berkeley.edu
{mhsu, zhihao.zhang}@haas.berkeley.edu

Abstract

We present results from a series of experiments that probe the ability of masked language models (MLMs), such as BERT and RoBERTa, to respond to general knowledge questions that do not have a single correct answer. Our investigation leverages the semantic fluency task from cognitive science, in which a variable number of exemplars from a semantic category (e.g., fruits) need to be produced in a specific order. It allows us to evaluate what MLMs know about common categories and their members, a representative type of one-to-many relational knowledge, and how they organize and query such knowledge. We developed incremental cloze tasks that reflect serial knowledge search, and show that MLMs, especially RoBERTa, are able to generate semantic fluency responses that strongly resemble responses from human subjects in both their content and dynamics. These findings contribute to the literature on whether and how masked language models can be used as knowledge bases, and also provide novel insights on their knowledge structure.

1 Introduction and Prior Work

Masked language models, such as BERT (Devlin et al., 2019), have risen to prominence in natural language processing (NLP) and have been successfully deployed a wide variety of linguistic tasks. Beyond these applications, the possibility that these models may also represent factual knowledge of the world has attracted increasing attention (Rogers et al., 2020; Kumar, 2021). To this end, recent studies have evaluated BERT (and other models) primarily on fill-in-the-blank factual questions, such as "Steve Jobs was born in [Mask]", and the metric of concern is the percentage of trials where the models produce the correct answer (Petroni et al., 2019; Jiang et al., 2020; Cao et al., 2021).

While the fill-in-the-blank question answering task considered in previous studies can be an effective way to examine certain types of factual knowledge that BERT captures, it represents only a subset of ways that general world knowledge can be used by humans (or algorithms). For example, when contemplating options of fast food chains for a meal, there are objectively many correct answers, and "correctness" is not necessarily the only goal (Zhang et al., 2021; Bhatia et al., 2021). While healthy humans access complex relational knowledge and find such answers effortlessly, it remains unclear how well masked language models are able to deal with more complicated and richly structured world knowledge for diverse tasks.

More formally, this type of open-ended world knowledge questions is captured by a so-called semantic fluency task widely used in cognitive psychology to examine knowledge retrieval in humans (Lucas et al., 1998; Binetti et al., 1995). In this task, participants are given a category (e.g., a type of reading material) and must provide as many examples of the category as possible. It has been used extensively in clinical settings to understand the cognitive characteristics of patients with intellectual disabilities (Nilsson et al., 2021), and is understood to engage a number of cognitive functions, including semantic knowledge retrieval (Oh et al., 2019).

The open-ended nature of semantic fluency also poses a novel challenge for an intelligent system in that its output needs to be organized in a certain way. For example, when asked to generate a list of animals, people's responses routinely prioritize more prototypical examples (e.g., cat) and exhibit serial dependence driven by semantic relatedness, such that cat and dog are grouped together more often than cat and duck (Gruenewald and Lockhead, 1980). Since masked language models have been applied to a myriad of downstream NLP tasks, understanding the structural properties of their outputs is increasingly important, as many real-world tasks, such as music recommendation, often involve multiple outputs where serial dependency

may or may not be desirable.

To address these questions, we propose a method of generating semantic fluency responses with BERT models, thereby extracting relational knowledge from them in a way that requires serial output and provides insight into their ability to replicate richly structured human results. In keeping with how performance of language models in standard NLP tasks is evaluated, we use human performance on the same questions as the benchmark and present comparisons that focus on distinct aspects of the output. We further compare the performances of BERT and RoBERTa throughout to examine whether and how pretraining procedures and parameters affect the humanlike nature of the outputs (Liu et al., 2019). To underscore the advantage afforded by contextualization of MLMs, we also compare their performances against a GloVe-based approach (Pennington et al., 2014).

2 Answering Semantic Fluency Questions with Language Models

2.1 BERT/RoBERTa

Our method poses semantic fluency as a series of repeated cloze tasks on incrementally modified versions of a sentence. Initially, a BERT (or RoBERTa, same below) model is given a sentence of the form "An example of [category] is a [MASK]." In subsequent cloze tasks, the sentence is modified to include the previous responses the BERT model gave. For example, if the category was "an electronic device", and BERT had already given the responses "watch", "laptop", and "phone" (in that order), the sentence for the next cloze task would be "Examples of an electronic device are [MASK], phone, laptop, and watch." Since BERT considers the bidirectional context of the MASK token, the method takes both the previous answers and the category into account when deciding on the next word to produce. Hence, our method seeks to reflect the serial, path-dependent nature of semantic fluency tests performed by humans, while retaining the basic form of cloze tasks commonly used in existing work on knowledge extraction from these models.

When selecting an output word for a particular cloze task, BERT uses the associated probability for each word that it assigns based on the cloze, and the selection is confined to the top 30 words. This probabilistic response production reflects the well-established finding that human memory search is probabilistic and can be modeled by a random walk in a semantic network (Abbott et al., 2012). An implication of our method's weighted random selection process is that different runs of the BERT model with the same prompt can result in vastly different output sequences, reflecting the fact that different human participants in a semantic fluency test can produce different word lists.

Additional constraints were applied to the output words that our method produces. First, a fixed length of the output list needs to be specified beforehand for each run, given the lack of wide agreement on the cognitive model for stopping during a fluency task. Second, we imposed that no duplicates are produced, including singularized or pluralized versions of previous responses, enforced with the Pattern library (De Smedt and Daelemans, 2012) Similarly, the category name or its variants was excluded from possible responses. Third, we restricted the words that BERT considers to be nouns and proper nouns, with NLTK's (Bird et al., 2009) part-of-speech tagging utility. Words that do not meet all of the aforementioned restrictions are removed from the options in the weighted random selection. If no words meet all of the criteria, the iterative process ends. While it is theoretically possible for the method to produce fewer words than the predetermined number, this did not occur in any of our experiments. In Section 5 of this paper, we report ablation studies that removed these constraints, and show that the performance of the models was virtually unaffected.

2.2 GloVe

In contrast to the BERT and RoBERTa models, GloVe does not consider the context in which a particular word is used. Our word generation process for GloVe mimics the one we use with BERT models, except that we start each list of words with a single word that captures the meaning of the category (e.g. "a weather phenomenon" starts with "weather"). We then consider the 30 words with the highest average cosine similarity to the words produced so far, including the start word. We then use these similarities as weights in a weighted random choice among the candidate words.

3 Semantic Fluency Data

3.1 Human Data

We collected human semantic fluency data[1] with Qualtrics on Prolific[2] (a crowdsourcing marketplace similar to Amazon Mechanical Turk) using three categories ("a fruit", "a type of car", "a weather phenomenon"). The categories were drawn from an influential study of category norms (Van Overschelde et al., 2004). Participants were asked to input up to 20 examples of the given categories into provided text boxes (with one answer per text box), and were instructed to use their own memory to answer the questions, without relying on any external sources of information. We did not impose a time limit upon the participants.

Spelling mistakes were corrected and all answers were singularized, and monikers (e.g., "chevy") were unified with formal names (e.g., "Chevrolet"). Additionally, we addressed a limitation of the off-the-shelf BERT models. Since BERT does not outputs multi-word responses, such as "blood orange", we converted multi-word variants of a particular answer to a one-word answer (e.g. "blood orange" becomes "orange"). These conversions did not affect semantically similar answers (e.g. “tangerine” and “clementine” were left alone). An example of a cleaned list of fruits from the human subjects was [apple, banana, orange, pear, plum, peach, jackfruit, watermelon, honeydew, mango, strawberry, blackberry, blueberry].

3.2 BERT/RoBERTa Data

Our BERT/RoBERTa data were produced from two off-the-shelf models in the HuggingFace Transformers libary (Wolf et al., 2020): BERT-Base (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019). No additional finetuning was performed on these pretrained models. We generated sequences from BERT to match the distribution of list lengths produced by the human subjects. The BERT data were subjected to the same cleaning procedure as the human outputs. An example of a list of fruits generated by RoBERTa was [pear, apple, plum, banana, raspberry, strawberry, grape orange, cherry, mango, pineapple].

3.3 GloVe Data

Our GloVe data were produced from the glove-wiki-gigaword-300 dataset (Pennington et al., 2014) from Gensim (Rehurek and Sojka, 2011).

4 Evaluations

4.1 Output Content

We first examined how well the content of the output, and in turn the category knowledge represented by BERT models, aligns with that of humans. To this end, we used the weighted Jaccard Score, a measure of the overlap between the outputs of the BERT-based models and the words produced by human subjects. For each category and model, we calculate the following score:

$$J = \frac{\sum_{w \in H \cap B} P(w)}{\sum_{w \in H \cup B} P(w)} \quad (1)$$

where H represents the set of all words produced by the humans, B represents the set of all words produced by the BERT-based model, and *P(w)* is the production rate of a given word *w* across all trials (from both Humans and BERT). As with the standard Jaccard Score, our score has a minimum of 0 and a maximum of 1.

The rationale for weighting words by their production rate is to more heavily penalize missing a common word (such as "apple", mentioned by 96.35% of human participants in the "fruit" category) than a rare word (like "date", mentioned by 1.56% of human participants). Thus, the weighted Jaccard Score measures the alignment of the two sets, weighted by the "importance" of words.

Across the three categories, the weighted Jaccard Score is consistently high, especially for the RoBERTa-Large model (Table 1). It also outperforms GloVe substantially in the type of car and weather phenomenon categories, while showing comparable performance for the fruit category. Overall, these results indicate that, in terms of the content of the knowledge about items belonging to these categories, the RoBERTa-Large model demonstrates a consistently strong performance well-aligned with the human subjects.

Interestingly, being the category with lower weighted Jaccard scores than the others for the two MLMs, the fruit category also had the longest

[1] Written informed consent was obtained from the participants, and the study protocol was approved by the Committee for Protection of Human Subjects at the authors’ institution. All participants received payment at the rate of $15/hour for their time. The full text of instructions given to the subjects is available at https://osf.io/bxhp7/?view_only=9b416db4e9f140119d8b3d61c8217e57

[2] N = 197, of which 119 identified as female. Participation was limited to those residing in the U.S., and the mean age was 33.5y.

	Fruit	Type of Car	Weather Phenomenon
BERT	.361 (.011)	.569 (.019)	.803 (.012)
RoBERTa	.672 (.013)	.785 (.015)	.795 (.013)
GloVe	.698 (.016)	.562 (.017)	.618 (.018)

Table 1: Weighted Jaccard Scores across models and categories (human vs. each language model). Standard errors in parentheses, obtained via a Bootstrap procedure with 10,000 iterations.

average list length (i.e., number of items produced by humans) at 13.0, compared with 8.0 for "type of car" and 7.3 for "a weather phenomenon". This may have contributed to the more divergent output content from humans and the models.

4.2 Output Dynamics

While the weighted Jaccard Score captures the alignment between the sets of words produced by the human subjects and the model, it does not take into account the dynamics of response generation, i.e., the order of words in the responses. More specifically, we sought to capture output structure by comparing the transitions between words that are exhibited by the human subjects and the models. To this end, we construct first-order adjacency matrices from our fluency data for each category (one for humans and one for each model). In the adjacency matrices, each row and column represents a unique word. Each entry of the matrix is the observed number of transitions from the row word to the column word in the sample. Therefore, this adjacency matrix quantitative captures the well-documented phenomena of serial dependence and clustering in semantic fluency.

To facilitate direct comparisons, each matrix has both its rows and columns specified by $H \cup B$, and both the rows and columns are kept in the same order in both matrices. As such, the matrices constructed from human and model responses can then be vectorized for computing the Pearson's Correlation Coefficient between them. As opposed to a transition probability matrix, the use of an adjacency matrix naturally gives more weight to the more frequently appearing items in the computation of the correlation coefficient.

Overall, from the similarity of the transitions and clustering patterns among the output items, RoBERTa demonstrates a consistent alignment with the dynamics demonstrated in human responses, and dominates the performance of the BERT-base model (Table 2), echoing the previous analysis that focuses on output content. Notably, as a non-contextualized benchmark, the GloVe-based approach show large deficiencies in producing human-like word transitions across all three categories, in sharp contrast with its moderate performance on output content. This finding highlights the advantage afforded by contextualization in BERT models in producing more nuanced transitional structure in their output.

	Fruit	Type of Car	Weather Phenomenon
BERT	.184	.166	.503
RoBERTa	.466	.442	.590
GloVe	.208	.099	.146

Table 2: Pearson's Correlation Coefficient values of vectorized adjacency matrices (human vs. BERT/RoBERTa/GloVe). 1 is the highest possible value. All p-values were < 0.001.

4.3 Holistic Discrimination: A Turing Test

Finally, we employ a Turing Test study with human participants[3] as a way to further confirm the ability of MLMs to produce humanlike responses for semantic fluency tasks as perceived by human observers. Since the RoBERTa-Large model consistently outperformed GloVE and BERT-Base, we only considered RoBERTa-Large in this experiment. We administered our Turing Test experiment on Prolific[4].

For this test, we removed any duplicate answers in both the human and RoBERTa-generated lists, and also ensured that participants were presented with human- and BERT-generated lists of the same length. This is to ensure that participants based their judgment on the identity and ordering of the items mentioned, corresponding to our analyses. Lists in the Turing test were randomly drawn from the previous data from humans and BERT, and only lists whose lengths were between 5 and 15 from both humans and BERT were used. In each trial, participants were asked to identify which of two given lists was generated by the BERT-based model. Each participant was given five pairs of lists per category.

[3]The full text of instructions given to the subjects is available at https://osf.io/bxhp7/?view_only=9b416db4e9f140119d8b3d61c8217e57

[4]U.S. subjects only. N = 146 for RoBERTa-Large with 90 females, mean age 30.4y. Study approval, the attainment of informed consent, and payment were the same as the semantic fluency studies before.

Across categories, RoBERTa-Large was able to fool the human judges to a substantial degree (Table 3). The proportion of incorrect responses in the Turing test was generally close to the chance level of 0.5, indicating that the semantic fluency responses generated with RobERTa-Large with our approach were nearly (albeit not perfectly) indistinguishable to human responses. In fact, a two-tailed z-test showed that the proportion of correct responses with all categories combined was not significantly different from the chance level ($p = 0.32$).

	Fruit	Type of Car	Weather Phenomenon
RoBERTa	.448 (.018)	.453 (.018)	.567 (.018)

Table 3: Proportions of incorrect answers in the Turing Test study across categories (higher number means more participants made incorrect choices). Standard errors in parentheses.

5 Ablation Studies

In this section, we explore the impact of removing some of the steps in our algorithm for generating semantic fluency outputs. In particular, we examine the outputs of BERT, RoBERTa, and GloVe on the same evaluation metrics for content (weighted Jaccard score) and dynamics (Pearson's correlation of the adjacency matrix with the one from human data) (1) when the part-of-speech (POS) constraint is removed and (2) when both the POS and duplicate constraints are removed. We find that both metrics remain largely similar across models and categories (Figure 1[5]). Therefore, our results on the performance of the MLMs and the comparison model with GloVe reflect intrinsic properties of these models and are unlikely to be contaminated by these post-processing steps.

6 Conclusion and Future Work

Taken together, our results show that probing BERT models with incremental cloze tests was an effective approach to extract one-to-many relational knowledge as prompted by the semantic fluency task, and that the responses generated were similar to human responses in both content and dynamics. Furthermore, the best performing MLM, RoBERTa, appears to be capable of organizing and outputting

[5]Numerical data and standard errors are available at https://osf.io/bxhp7/?view_only=9b416db4e9f140119d8b3d61c8217e57

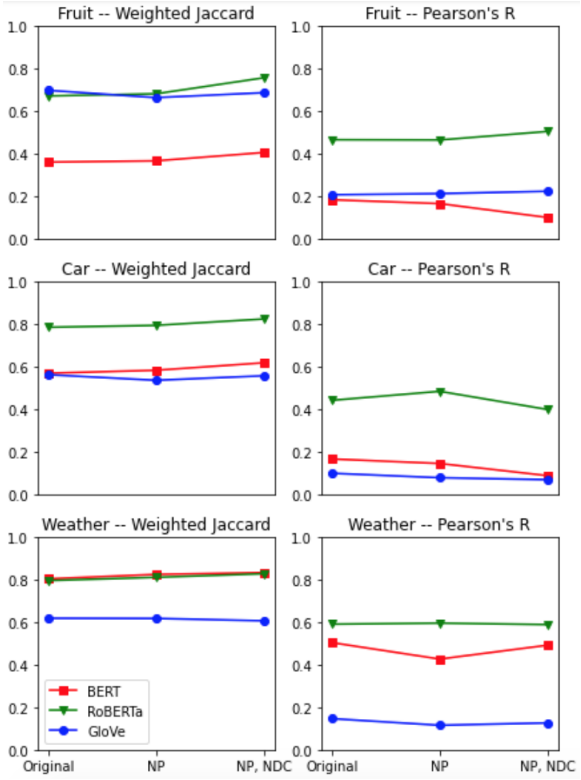

Figure 1: Weighted Jaccard Scores and Pearson's R for original and ablated models. NP is No POS tagging. NP, NDC is No POS Tagging and no duplicate constraint.

knowledge in a way that could be fairly humanlike (albeit not yet perfectly) to the eyes of human judges, as shown by the Turing Test results. In addition, RoBERTa robustly outperforms BERT across evaluation metrics, providing additional support for the importance of training optimization with a novel task. While our experiments do not provide a full picture of the degree to which language models represent knowledge in the same manner as humans, our work broadens how language models may be probed and used as knowledge bases (Rogers et al., 2020; Petroni et al., 2019), extending extant work to deal with more richly structured outputs in response to inquiries of general world knowledge.

Our experiments were limited by the fact that the BERT models could only produce single-word responses. Adapting a BERT model to perform the semantic fluency task with multi-word responses could lead to explorations of BERT's ability to produce richly structured humanlike responses in a wider range of categories. Additionally, our prompt structure only represents one way of using repeated cloze tasks to elicit complex information from BERT models. Hence, variations on our

prompting method can be constructed to produce semantic fluency responses. Finally, our Turing Test does not differentiate between the impacts of the content and order of words in responses in a human judge's verdict. Alternative tests may provide insight into the extent to which human judges use content and order to make their decisions.

References

Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. 2012. Human memory search as a random walk in a semantic network. In *NIPS*, pages 3050–3058.

Sudeep Bhatia, Lisheng He, Wenjia Joyce Zhao, and Pantelis P Analytis. 2021. Cognitive models of optimal sequential search with recall. *Cognition*, 210:104595.

Giuliano Binetti, Eugenio Magni, Stefano F Cappa, Alessandro Padovani, Angelo Bianchetti, and Marco Trabucchi. 1995. Semantic memory in alzheimer's disease: An analysis of category fluency. *Journal of Clinical and Experimental Neuropsychology*, 17(1):82–89.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul J Gruenewald and Gregory R Lockhead. 1980. The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, 6(3):225.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Abhilasha A Kumar. 2021. Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28(1):40–80.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John A Lucas, Robert J Ivnik, Glenn E Smith, Daryl L Bohac, Eric G Tangalos, Neill R Graff-Radford, and Ronald C Petersen. 1998. Mayo's older americans normative studies: category fluency norms. *Journal of clinical and experimental neuropsychology*, 20(2):194–200.

Karin Nilsson, Lisa Palmqvist, Magnus Ivarsson, Anna Levén, Henrik Danielsson, Marie Annell, Daniel Schöld, and Michaela Socher. 2021. Structural differences of the semantic network in adolescents with intellectual disability. *Big Data and Cognitive Computing*, 5(2):25.

Se Jin Oh, Jee Eun Sung, Su Jin Choi, and Jee Hyang Jeong. 2019. Clustering and switching patterns in semantic fluency and their relationship to working memory in mild cognitive impairment. *Dementia and neurocognitive disorders*, 18(2):47–61.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

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

James P Van Overschelde, Katherine A Rawson, and John Dunlosky. 2004. Category norms: An updated and expanded version of the norms. *Journal of memory and language*, 50(3):289–335.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhihao Zhang, Shichun Wang, Maxwell Good, Siyana Hristova, Andrew S Kayser, and Ming Hsu. 2021. Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proceedings of the National Academy of Sciences*, 118(20).