Give or Take a Few Representations

John R. Starr, Jacob Matthews

Over the past few years, transformer models (BERT, GPT2, RoBERTa, among others) have achieved state-of-the-art performance on a plethora of NLP tasks. A primary reason for the success of transformer models is that they utilize *contextual embeddings*, where the representation for each word is based on the specific context it appears in; contextual embeddings differ from *static embeddings*, where each word is assigned one specific representation. Contextual embeddings help the model disambiguate different senses of a word, given unique contexts ("river bank" vs. "money bank").

In this report, we investigate how contextual embeddings from large-scale transformer models represent *ambiguous phrasal verbs* (APVs), shown below in (1) and (2):

(1) The teacher had given in to the student's request.

(2) The assignments the teacher had given in class were too long!

In (1), we find a *true phrasal verb*: "given" and "in" are acting as a constituent. In (2), we find a *false phrasal verb*: "given" and "in" are not a constituent, instead "in" is the head of the prepositional-phrase complement for "given".

Previous work on APVs has found that linguistic features (dependency head, lexical item) can be used to classify an APV as either a true phrasal verb or a false phrasal verb (Tu & Roth 2011, 2012); these findings were supported by later work that performed the classification task with neural networks (including transformer models) and their embedding representations (Shwartz & Dagan 2019). However, very little is known about *how* transformer models represent APVs. We fill this gap by extracting embeddings for APVs from three transformer models – BERT (Devlin et al. 2018), GPT2 (Radford et al. 2017), and RoBERTa (Liu et al. 2019) – across five layers, then performing a representational analysis using both within-model and across-model comparisons. While all models perform significantly better than a majority baseline on the classification task, the method and representations they use to do so appear to differ. All code, data, and analyses for this report can be found on the first author's GitHub page[1].

[1] https://github.com/johnstarr-ling/light-verb-construction-embeddings

1: DATASET AND METHODOLOGY

For this project, we utilize the PVC Data corpus constructed for Tu & Roth (2012), which is a subset of the British National Corpus (BNC). The corpus contains 1348 sentences, each with an APV from a set of 23 possible APVs. Each sentence has annotations for the APV's status as either a true phrasal verb or a false phrasal verb, the inter-annotator agreement on the phrasal status of the APV, and whether the APV is "idiomatic" or "compositional". We removed all sentences that had a low level of inter-annotator agreement (<80%). Therefore, the final dataset for this project consisted of 1224 total sentences (~65% true phrasal verbs, ~35% false phrasal verbs).

For each sentence's APV, we extracted the contextual embeddings from three transformer models (BERT, GPT2, and RoBERTa) at five layers [0, 3, 6, 9, 12] using `minicons` (Misra 2022). Embeddings from these layers were extracted to unveil how the trajectory of APV embeddings may change over time within and across models. Additionally, we extracted the embeddings for both the verbal element in the APV ("given" in "given in") and for the non-verbal element in the APV ("in" in "given in") as a point of comparison to the possible effects of mean pooling on creation of phrasal embeddings. In total, these contextual embeddings serve as the input to a classifier that determines whether a sentence is a true phrasal verb or a false phrasal verb (Section 2), along with being the primary focus of the representational analysis (Section 3).

2: CLASSIFICATION TASK

As mentioned in the introduction, previous work has established that featural representations (either using explicit linguistic annotations or the final layer of learned neural embeddings) are capable of distinguishing the phrasal status of APVs (Tu & Roth 2011, 2012; Shwartz & Dagan 2019). We aim to replicate these findings, while expanding prior analyses by checking performance across layers and embedding types. For each type of embedding from each model, we train a 5-fold cross-validated SVM classifier to predict whether an APV is a true phrasal verb or a false phrasal verb. Given the dataset consists of approximately 65% phrasal verbs, we consider the majority baseline to be 0.65 accuracy. The results can be found below in Figure 1:

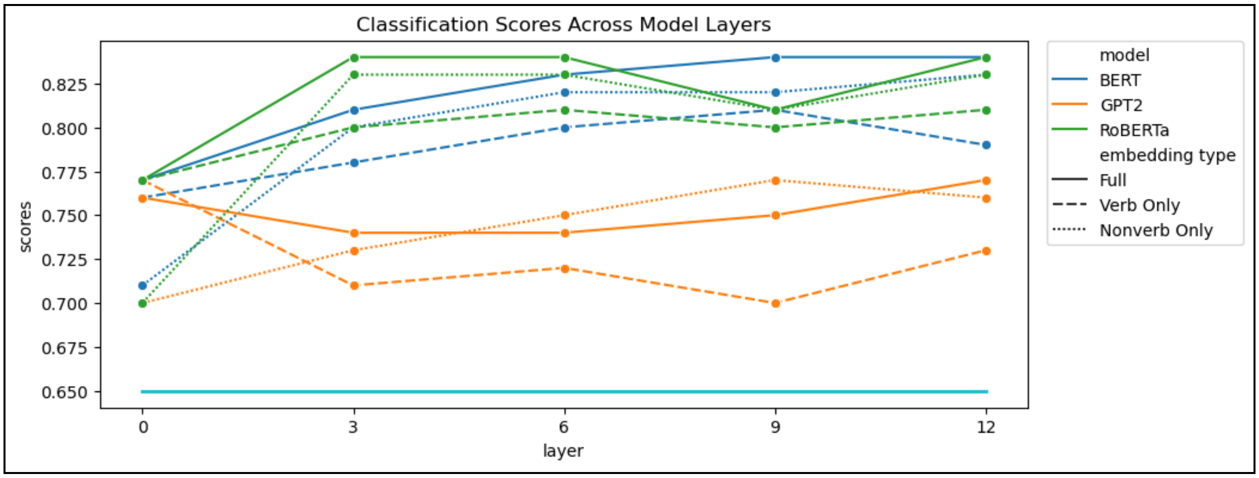

Figure 1: Classification accuracy across model layers for each model, by embedding type. Horizontal teal bar indicates majority baseline.

We replicate previous findings in showing that contextual representations can be used to predict the phrasal status of an APV more accurately than selecting the majority class, suggesting that the contextual embeddings encode specific distinctions between true phrasal verbs and false phrasal verbs. In particular, we find that BERT and RoBERTa better predict the phrasal status of APVs more accurately than GPT2. Additionally, we report differences in how each model's representations affect classification accuracy across layers: BERT embeddings appear to improve incrementally with each layer, while RoBERTa improves rapidly between layers 0 and 3 and maintains that level (with a slight dip at layer 9) through the final layer. However, GPT2 only shows improvement across layers when using the non-verb embeddings only; both full verb and verb only embeddings do not improve across layers. To understand what might be driving these differences between models, we perform three representational analyses in the following section.

3. REPRESENTATIONAL ANALYSES

In this section, we conduct three representational analyses, each examining one of the following questions respectively: a) how do model representations of these APVs change over time? b) how do different embedding types (full verb, verb only, non-verb only) compare to one another across layers within each model? and c) are the relationships between representations that each model creates similar to one another, or does each model create unique relationships between representations?

3.1 *Model Representations Diversify Across Layers*

To visualize how model representations vary across layers, we use the t-SNE method to reduce the dimensionality of each APV's full embedding to two dimensions for layers 0, 6, and 12; these two dimensions are arbitrary and are uninterpretable. For results, see Figures 2-4 below, where color indicates different APVs:

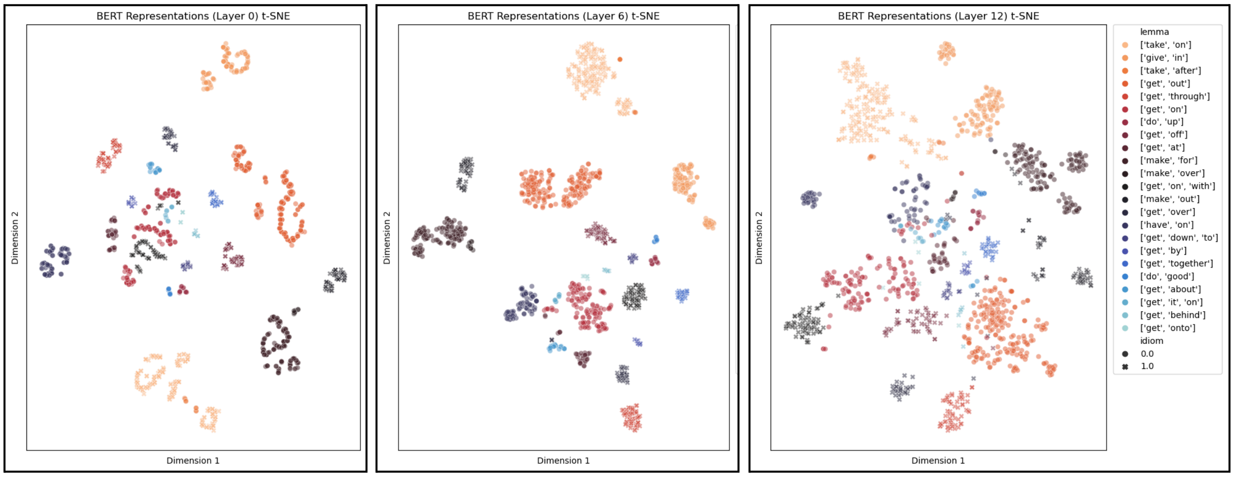

Figure 2: t-SNE of BERT representations from layers 0 (left), 6 (center), and 12 (right).

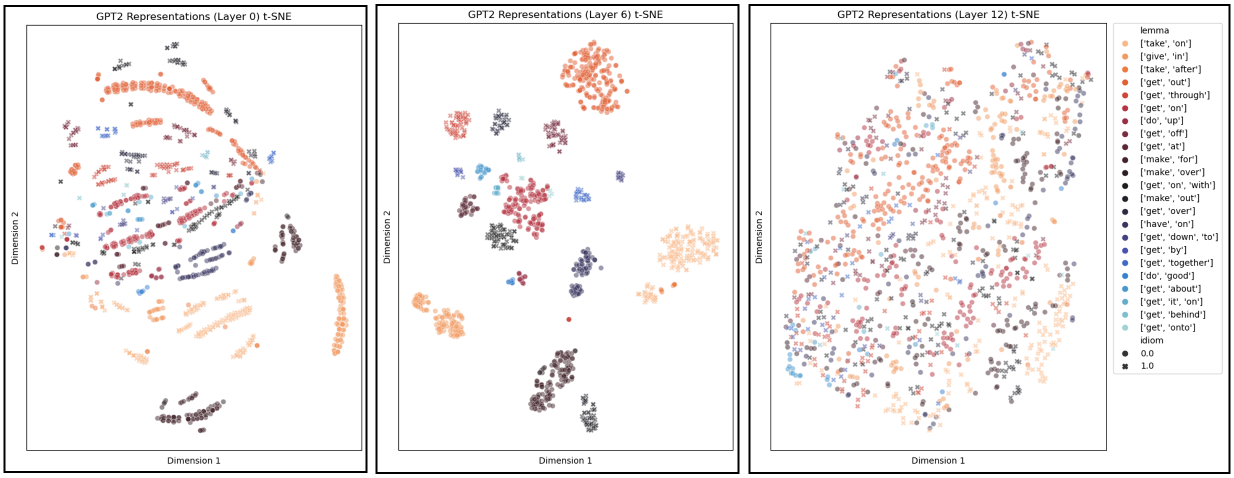

Figure 3: t-SNE of GPT2 representations from layers 0 (left), 6 (center), and 12 (right).

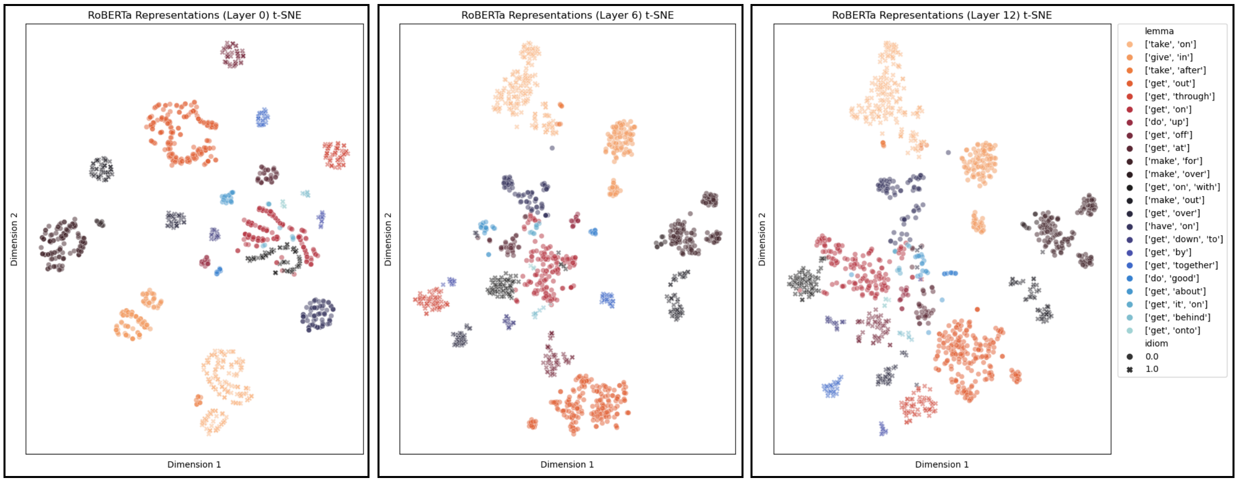

Figure 4: t-SNE of RoBERTa representations from layers 0 (left), 6 (center), and 12 (right).

Across models, we find that APVs cluster in small, bacteria-like clusters in layer 0; these clusters appear to be distinct for each APV. In layer 6, all models maintain strong semblances of clusters, though noise in the embedding space increases (especially for RoBERTa). However, in layer 12, we note a clear break from the observed patterns from layers 0 and 6: BERT and RoBERTa maintain clusters in the representational space, while GPT2 appears random.

It is currently unclear how these findings are reconcilable with the performance scores on the classification task: while GPT2 appears to dramatically change its representations in the final layer, we do not find a significant change in performance in the classification task for the final layer of GPT2. We will discuss reasons why we find these results in the following sub-section.

3.2 *Similarity Between Embedding Types Decreases Across Layers*

To better understand how each model creates contextual embeddings, we compare the cosine similarity between embedding types within each model, with three possible comparisons: {full:verb embeddings, full:non-verb embeddings, verb:non-verb embeddings}. Additionally, we perform this calculation within each layer to see how similarity changes over time. For example, given the APV "given in", we compute the cosine similarity for the following embeddings: {"given in":"given", "given in":"in", "given":"in"}. Results for this analysis can be found below in Figure 5:

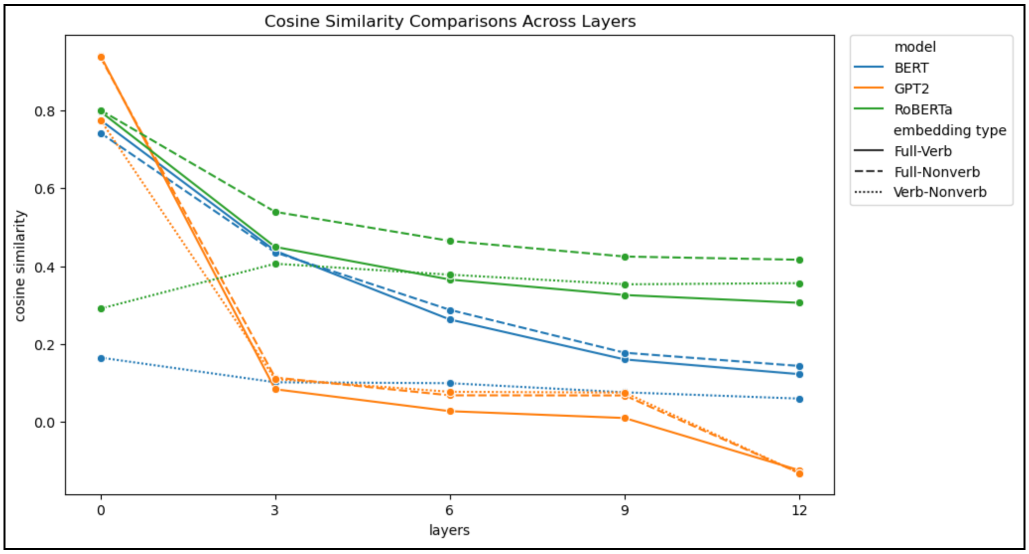

Figure 5: Cosine similarity between embedding types within each model.

Generally, we find that cosine similarity decreases over time across all comparisons and across all models, suggesting that the contextual embeddings that the models are learning are becoming more dissimilar as we move to higher layers of the model. This is intuitive: as the model increases degrees of abstraction, the representations in these spaces are more likely to spread further apart. However, the rate of separation strongly differs across models: BERT and RoBERTa incrementally decline over time, while GPT2 rapidly declines between layers 0 and 3, stabilizes, then drops significantly in the final layer. In short, GPT2 appears to learn completely unique representations for each word in the final layer, though this dramatic shift is not present prior. Such dramatic changes in representation in the final layer provide some explanation for the findings in the previous subsection: GPT2 learns unique representations for each APV, meaning the representational space will appear noisy.

We also note distinctions in the processing of verb-nonverb comparisons (dotted lines in Figure 5) across models. For both BERT and RoBERTa, cosine similarity between verb and nonverb elements of an APV does not appear to change across layers. In GPT2, verb-nonverb similarity analyses identically follow the patterns observed for full-verb and full-nonverb comparisons. We posit that the differences in the earlier layers between GPT2 in relation to BERT and RoBERTa for the verb-nonverb comparison may stem from the auto-regressive nature of GPT2: since the non-verb appears right after the verb, the context does not dramatically change. Models such as BERT and RoBERTa, which are not auto-regressive, have contextual information both before and after the verb and non-verb elements.

3.3 *Relationships Between Representations are Similar Across Models*

The two previous representational analyses suggest that the representations for APVs are more stable across layers in BERT and RoBERTa, whereas GPT2 shows significant variation. Yet, it is still unclear if the relationships between representations in one model are distinct from the relationships between representations in another, or if they construct similar relationships between representations. To test this, we calculate the cosine distances for each APV against all other APVs for each model. Then, we compute the cosine similarity of these distances across models: we call this *second-order cosine similarity*.

For a more concrete example of how we compute second-order cosine similarity, consider the following two APVs: “given in” and “took on”. First, we determine how similar the representations for “given in” and “took on” are for both BERT and for GPT2 using cosine distance. This will produce a 2x2 matrix for both BERT and GPT2 of how similar “given in” and “took on” are within the model. Then, we calculate the cosine similarity of each row in these 2x2 matrices between the BERT and GPT2 representations. These comparisons will tell us how comparable the relationship between “given in” and “took on” is between BERT and GPT2 . Additionally, we repeat this process to the embeddings from layers [0, 3, 6, 9, 12]. The results for these second-order cosine similarity calculations can be found in Figure 6 below:

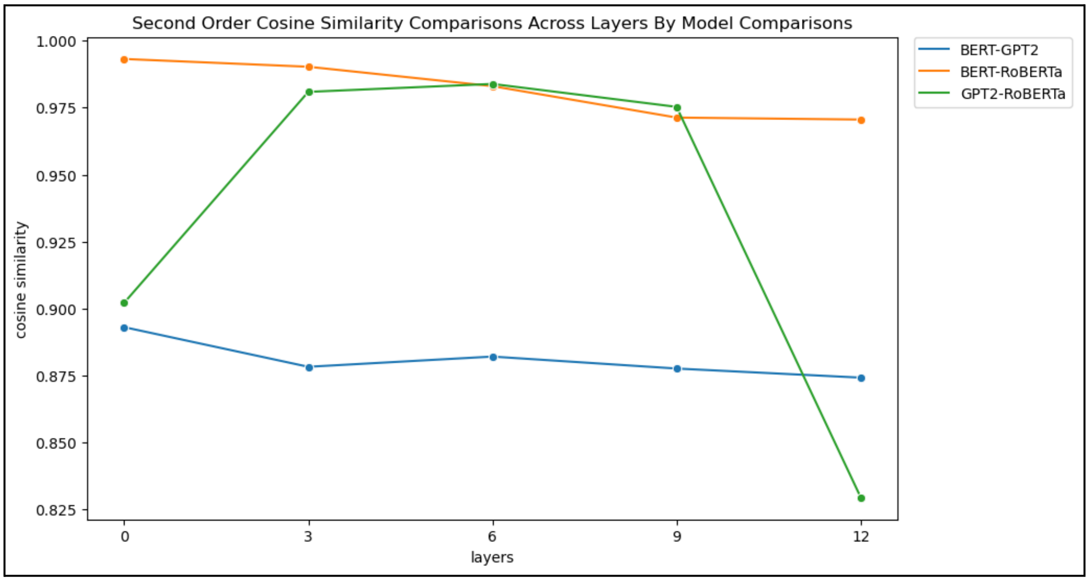

Figure 6: Comparison of the relationships between representations within each model, by layer (mean second-order cosine similarity). 95% CIs not shown, as all CIs < 0.0005.

We find that all models construct relationships between APV representations with a high degree of similarity (>0.825). However, we note that some model comparisons show distinct patterns of behavior. For example, BERT and RoBERTa construct similar relationships of representations across all layers, supporting our findings in the previous representational analyses that show the two models tend to behave similarly. This similarity likely stems from the models having nearly-identical architectures (with modified training procedures). Additionally, we report that GPT2 and RoBERTa produce significantly different relationships between representations depending on the layer. Layers 3, 6, and 9 show the highest similarity, but the final layer shows a significant decrease in similarity for how the models create relationships between representations.

4. DISCUSSION

In the classification task, we replicate previous work on the topic, finding that contextualized representations strongly predict the phrasal status of APVs.We contribute to previous research by including how performance changes across layers, finding that both BERT and RoBERTa increase over time, while GPT2 does not change.

Generally, our representational analyses confirm the findings of the classification task: BERT and RoBERTa represent APVs in a similar manner, whereas GPT2 displays greater variation across layers. In particular, we note a dramatic change in representations within the last layer of GPT2: the model appears to have completely unique representations for each individual instance of an APV, even if it found similarities across APVs prior. These analyses suggest that differences between models may stem from their distinct architectures (auto-regressive vs. not auto-regressive, encoder only vs. decoder only) or their parameter sizes (GPT2 > BERT, RoBERTa), though it is currently unclear as to what exactly is driving the representational differences.

In total, these findings contribute to the probing literature within NLP and computational linguistics by unpacking representational differences across layers for three state-of-the-art transformer models. Future research on this topic will incorporate additional models, more across- and within-model comparisons, along with extending this work to different kinds of ambiguous representations, such as *idioms* or *metaphors*.

References

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*.

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*.

Misra, K. (2022). minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Shwartz, V., & Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, *7*, 403-419.

Tu, Y., & Roth, D. (2011, June). Learning English light verb constructions: contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. 31-39.

Tu, Y., & Roth, D. (2012). Sorting out the most confusing English phrasal verbs. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. 65-69.