Using Paraphrases to Study Properties of Contextual Embeddings

Laura Burdick, Jonathan K. Kummerfeld and **Rada Mihalcea**
Computer Science & Engineering
University of Michigan, Ann Arbor
{lburdick,jkummerf,mihalcea}@umich.edu

Abstract

We use paraphrases as a unique source of data to analyze contextualized embeddings, with a particular focus on BERT. Because paraphrases naturally encode consistent word and phrase semantics, they provide a unique lens for investigating properties of embeddings. Using the Paraphrase Database's alignments, we study words within paraphrases as well as phrase representations. We find that contextual embeddings effectively handle polysemous words, but give synonyms surprisingly different representations in many cases. We confirm previous findings that BERT is sensitive to word order, but find slightly different patterns than prior work in terms of the level of contextualization across BERT's layers.

1 Introduction

Contextualized embedding algorithms, such as BERT (Devlin et al., 2019), have achieved impressive performance on a wide variety of tasks (Huang et al., 2019; Chan and Fan, 2019; Yoosuf and Yang, 2019). One application of BERT is using it as a measure of sentence similarity (Zhang* et al., 2020; Sellam et al., 2020), based on the assumption that BERT will produce similar representations for the words in two sentences with similar semantics.

We propose to use paraphrases with alignments between words as a tool for studying how BERT represents words and phrases. Figure 1 shows an example. Critically, when considering an aligned word pair, we can assume the context has a similar impact on both words because we know the phrases are semantically similar. Previously, paraphrases have been used to probe whether compositionality is accurately captured by BERT (Yu and Ettinger, 2020), but we believe they can be used to explore many other questions.

Using the second version of the Paraphrase Database (PPDB, Pavlick et al., 2015), we explore how consistent contextual representations are when controlling for the semantics of the context. First, we use the human-annotated portion of the PPDB data to confirm that BERT consistently represents paraphrases. Next, we use the highest-quality (but not all human-annotated) section of the PPDB to probe BERT's behavior in more detail. Looking at words, BERT effectively handles variations in spelling, but does less well with spelling errors. BERT also effectively handles words of varying levels of polysemy, but the representations for synonyms are surprisingly diverse, with a much broader distribution of similarity scores. These findings confirm results from prior work using other methods, while uncovering new insights about contextual embedding models.

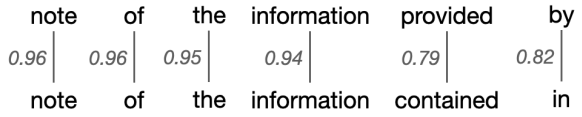

Figure 1: Example paraphrase from the PPDB with word alignment and word cosine similarities using the last layer of BERT.

We also consider a range of other models' word representations, finding that they have similar patterns to BERT, but with aligned words that are the same receiving even more consistent representations than from BERT. BERT gives less contextualized representations to paraphrased words than non-paraphrased words, with the exception of punctuation. Finally, we re-evaluate work looking at patterns across BERT's layers and find that when controlling for semantics, the later layers actually produce more similar representations (in contrast to previous work).

These results show that paraphrases are a useful tool for studying representations. By controlling

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4558 - 4568

for meaning while presenting interesting surface variations, they provide a unique probe of behavior.

2 Background

2.1 BERTology

There has been a growing body of research studying the inner workings of BERT and trying to quantify what it learns in various scenarios, dubbed "BERTology" (Rogers et al., 2020). Of particular interest to this paper is work that analyzes BERT's output embeddings. Recent studies have found that embeddings created from the final layer of BERT tend to cluster according to word senses (Wiedemann et al., 2019), though this varies somewhat based on the position of a word in a sentence (Mickus et al., 2020). The final BERT layers also produce more contextualized word embeddings than the earlier layers (Ethayarajh, 2019), a finding we revisit using paraphrases in Section 3.3.

2.2 The Paraphrase Database

To analyze BERT, we take advantage of the unique properties of paraphrases. We use the Paraphrase Database (PPDB, Ganitkevitch et al., 2013; Pavlick et al., 2015), a database of paraphrases collected using bilingual pivoting, the process of taking a particular English phrase, looking at all the foreign language phrases it can be translated into, finding all occurrences of these foreign language phrases, and then translating them back into English (Bannard and Callison-Burch, 2005). PPDB 2.0 contains 100m+ English paraphrases, each with word alignment information, an automatically generated quality rating, and, for a subset, a human quality rating.[1] Word alignments are the by-product of the bilingual pivoting method used to collect the paraphrases. When using alignments, we only consider phrases from the highest quality section of the PPDB, which are most likely to have accurate alignments. Example paraphrases with their average human annotations and automatically generated scores are shown in Table 1. In general, the phrases in this dataset are short. The longest phrases have six tokens, and the majority have fewer than six.

Human quality ratings are included for 26,455 paraphrase pairs, with five annotations per paraphrase. Agreement is measured using Spearman's ρ (Spearman, 1910); the average ρ between two workers is 0.57, and the average ρ between each worker with the other four annotators is 0.65.

[1] http://paraphrase.org.

Phrases	Human Score	PPDB Score
are you talking do n't they	1.0	2.7
what 's this all about ? what 's she saying ?	4.2	3.9
where did they come from ? where are they from ?	4.8	4.4

Table 1: Example tokenized paraphrases from the PPDB, with their average human annotations and automatic PPDB scores.

The automatic quality ratings (PPDB score) are generated by using the human annotations to fit a supervised ridge regression model. The input to the model consists of 209 hand-crafted paraphrase features, including WordNet features (Fellbaum, 1998), distributional similarity features, and cosine similarities of generated Multiview Latent Semantic Analysis embeddings (Rastogi et al., 2015). The PPDB score achieves a Spearman's ρ of 0.71. In comparison, Pavlick et al. (2015) report that using the word2vec embedding of the rarest word in each paraphrase obtains Spearman's ρ of 0.46.

3 Experiments

In our experiments, we want to use the PPDB to examine BERT's ability to consistently represent paraphrase semantics.[2] In order to do this, we consider both phrase-level and word-level embeddings. Except where explicitly indicated otherwise, all experiments are run using the uncased base model of BERT, using a maximum sequence length of 128 and a batch size of 8. We use the pretrained models provided by the Transformers library.[3]

There is a slight mismatch between the PPDB's tokenization and the format of the BERT training data. The mismatch primarily occurs with contractions and apostrophes (e.g., BERT expects "don't", while the PPDB is tokenized "do n't"). This does not substantially affect the results; less than 8% of the human-annotated paraphrase pairs contain apostrophes. When words are broken into multiple pieces by the wordpiece tokenizer, we use the

[2] Note, paraphrases do not always have identical meaning. We focus on particularly similar pairs for our analysis to support our assumption that their meaning matches.

[3] https://huggingface.co/docs/transformers/index

average of the pieces as the word representation.

3.1 Phrase-Level Embeddings

First, we consider phrase-level embeddings that capture aggregate information about all of the words in a given phrase. These embeddings show us that BERT is able to distinguish between two paraphrases, and two unrelated phrases.

We use 25,736 phrase pairs with human annotations in the PPDB.[4] Each human annotation is between 1 and 5, reflecting the similarity of the two phrases. We run each phrase through the pretrained BERT model. For each pair of phrases, we average together the embeddings for each word to get a phrase embedding. We create phrase embeddings using averaging because previous research has shown that this method is effective. For example, Reimers and Gurevych (2019) created sentence embeddings using three methods: (1) averaging word embeddings, (2) taking the maximum of word embeddings, and (3) using the CLS token vector. They found that averaging created the best sentence embeddings for semantic textual similarity tasks.

After creating phrase embeddings, we take the cosine similarity between the two embeddings. We compare with ground truth annotations using Spearman's ρ. We do this for each of the twelve BERT layers, and the concatenation of all layers. We use cosine similarity to compare embeddings because this metric is commonly used when working with BERT (e.g., Mahmoud and Torki (2020); Garí Soler and Apidianaki (2020); Kovaleva et al. (2019)).

We compare BERT to a more traditional embedding method, the continuous bag-of-words approach in word2vec (w2v) (Mikolov et al., 2013). We train w2v on an English Wikipedia corpus of 5,269,686 sentences,[5] using dimension size 200, a window size of five, and a minimum count of five. We choose to train w2v on Wikipedia data, in order to replicate the correlations in Pavlick et al. (2015). We train five w2v models, using five different random seeds.[6] For each pair of phrases, we average together the embeddings for each word to get a phrase embedding, and then take the cosine similarity between the two phrase embeddings.[7] We report the average and standard deviation of Spearman's ρ over the five models.

Comparing Sentences and Phrases One difference between our work and the way BERT is normally used is that we have phrases rather than sentences. To check that this does not substantially change BERT's behavior, we compare the embeddings for phrases in a sentence and the phrases on their own. We take 9,780 paraphrases from the PPDB. We choose paraphrases where one of the phrases has at least six tokens, the paraphrase has a relatively good PPDB score and no syntactic placeholders. This is described further in Section 3.2. For each phrase, we find up to 100 sentences (on average, 80.5 sentences) in Gigaword (Parker et al., 2011; Rush et al., 2015)[8] and OpenSubtitles (Tiedemann, 2012)[9] that contain that phrase. For each sentence, we run it through BERT and average together the word embeddings for words in the phrase to create a phrase embedding. The phrase embeddings are very similar across different sentences (average cosine similarity of 0.82 ± 0.07).

Now we can compare (1) the average of phrase embeddings derived from sentences, with (2) embeddings for phrases in isolation, to see if BERT will be confused by not having a complete sentence. For each phrase, we take the cosine similarity between the phrase embeddings created using these two methods. The phrase embeddings are fairly similar (average similarity of 0.74 ± 0.12). This gives us confidence that BERT produces embeddings for phrases on their own that are very similar to phrases in the context of a sentence. For the rest of our experiments, we run phrases individually through BERT, rather than in the context of complete sentences, which allows us to focus on the semantics of the phrase itself.

Results on the PPDB. Table 3 shows results for BERT, w2v, and the PPDB model, broken down by the average length of each paraphrase. For all layers, BERT improves on longer paraphrases. This is intuitive, because the longer the phrase, the more it will be able to leverage contextual information. The last layer of BERT behaves slightly differently than the other layers. While it continues to perform

[4] This is 3% smaller than the entire human-annotated subset. We were unable to map some of the human-annotated data to the data with PPDB scores (even with help from the authors of the PPDB paper). This may be why our scores for w2v are lower than those reported by Pavlick et al. (2015).

[5] This data was used in Tsvetkov et al. (2016) and is available by contacting the authors of that paper.

[6] 2518, 2548, 2590, 29, 401

[7] For both BERT and w2v, we additionally tried using the embedding of only the rarest word (with frequency measured using the full PPDB), as reported in Pavlick et al. (Pavlick et al., 2015), but this gave us consistently lower correlations.

[8] https://huggingface.co/datasets/gigaword.

[9] https://opus.nlpl.eu/OpenSubtitles.php.

Category	Phrases		BERT-sim
High Sim	the transport and illegal detention of	the transportation and illegal detention of	0.99
	representative of the secretary-general on	special representative of the secretary-general on	0.98
Low Sim	the ohchr	the high commissioner for human rights	0.32
	so , why	- does your shoulder bother you	0.42
Idiom	are mad .	're out of your mind .	0.67
	is everything all right , sir	are you okay	0.79

Table 2: Examples of common phenomena observed in paraphrases with particularly high similarity, low similarity, and idiomatic expressions.

	Average Length			All
Method	1-2.5	2.5-4	4-6	
BERT	0.2	0.4	0.51	0.31
w2v Average	0.35	0.32	0.41	0.43
PPDB model	0.41	0.50	0.51	0.50
Num. phrases	17,517	5,349	2,870	25,736
Avg. human	2.40	2.94	3.26	2.60

Table 3: Spearman's ρ between human-annotated PPDB paraphrases and different embedding methods (BERT and w2v), broken down by average paraphrase length (the average number of words in each of the two phrases in the paraphrase). Annotations are a score between 1 and 5. At the bottom of the table, we include the length distribution of the human-annotated paraphrases, as well as the average human annotation for each set of grouped lengths. For all length groupings, the w2v std. dev. is 0.0. For paraphrases length 1-2.5, the avg. human std. dev. is 1.0; for all other groupings, the std. dev. is 1.1.

better on longer paraphrases, it does substantially worse on short paraphrases and slightly worse on medium paraphrases.

Similarly, w2v also improves on longer paraphrases. By taking the average of all the word embeddings for each word in the phrase, w2v has more information to incorporate into its phrase embeddings for longer paraphrases. Though w2v improves as the paraphrases grow longer, it underperforms BERT for all but the shortest paraphrases. We also see that the automatic PPDB score does better on longer paraphrases. This could be because it incorporates distributional information, which is richer when there are more words. Finally, the human annotation scores show that longer paraphrases are more similar.

From Table 3, we see that the final layer of BERT outperforms w2v and performs comparably to the PPDB score on the longest paraphrases. This is not a completely fair comparison; the PPDB model is trained specifically on this data, and has access to outside information that BERT does not, such as WordNet features and additional features derived from the translation process used to create the PPDB. These results give us confidence that BERT can distinguish between phrases that are paraphrases of each other and phrases that are not.

Looking at BERT's output, we can see several patterns in the paraphrases that receive high and low similarities. Table 2 shows examples of these patterns. For phrases with high similarity according to BERT, a single word changes or a single word is added. On the surface, these changes have very little impact on the meaning, though the addition of the word 'special' in the second case could change who is being referred to. For phrases with low similarity according to BERT, they frequently required world knowledge (e.g., definition of an acronym) or appeared to be errors. We also observed idioms getting reasonably high scores, but not as high as the literal paraphrases.

Conclusion: The standard way of using BERT to produce a representation of a phrase is consistent with human scores of paraphrases. All layers are effective, though the last layer struggles with shorter phrases.

3.1.1 One-Word Paraphrases

In Section 3.1, we saw that BERT does not do as well on short phrases as it does on longer ones. We explore the extreme case of single word paraphrases here. Among the subset of one-word paraphrases, there is a wide range of human annotations (average annotation 2.27 ± 0.99). To explore this further, we focus on one-word paraphrases with a human annotation of 5, the highest annotation score, indicating that these are the strongest synonyms. Among these high-quality synonyms, cosine similarities are consistently high for the last layer of BERT (average similarity 0.76 ± 0.12).

Table 4 shows synonyms with both the highest and lowest BERT similarities. Misspelled

Phrase 1	Phrase 2	Cos. Sim.
laboratoires	laboratories	0.51
completly	totally	0.51
fervor	enthusiasm	0.52
79.0	seventy-nine	0.53
approximatly	around	0.54
-mom	-mother	0.91
1.350	1.35	0.92
characterises	characterizes	0.92
km	kilometres	0.92
garbage	trash	0.96

Table 4: Cosine similarity scores for the last layer of BERT for one-word paraphrases with the highest human annotation score.

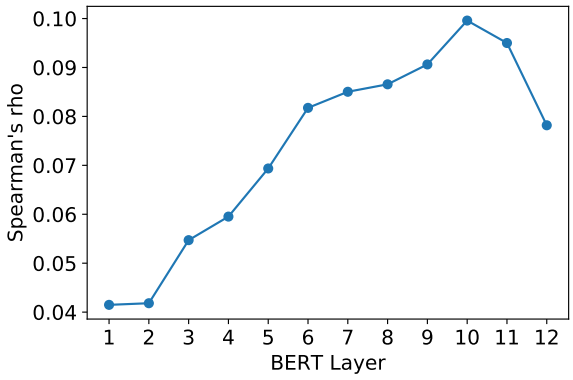

Figure 2: Spearman's ρ between BERT cosine similarities and PPDB scores for all aligned same words, broken down by BERT layer.

words (e.g., `completly`, `approximatly`) and pairs that involve different languages (e.g., French `laboratoires`) have low cosine similarities. Numbers appear on both the low end (e.g., `79.0` and `seventy-nine`) and the high end (e.g., `1.350` and `1.35`) of the similarity spectrum. One difference between the similar and dissimilar number pairs is that in the similar case they both use digits, while in the dissimilar case, one uses digits while the other uses words.

Conclusion: Looking at single words shows that BERT struggles to identify synonyms, and does particularly poorly with misspellings and cross-lingual comparisons.

3.2 Word-Level Embeddings

PPDB provides alignments between words in the paraphrases, automatically generated as part of bilingual pivoting. We use these alignments to consider four different sets of words:

Same, Aligned Words that are the same in both phrases and aligned.

Same, Unaligned Words that are the same in both phrases, but not aligned. These tend to be function words. 90% of our examples are one of (the, of, ", to, i, in, that, as, what). This category may have more examples of other word types if longer paraphrases are considered in future work.

Different, Aligned Words that are aligned, but not the same. This case covers synonyms.

Different, Unaligned Words that are not aligned and not the same (but still one from each phrase in a paraphrase pair). Note, these words are not completely unrelated. They are from the same paraphrase, making them more related than words from two random phrases.

In this section, we go beyond the human-annotated data considered in the previous section. We restrict our experiments to the highest quality paraphrases in the PPDB: dataset S. We also only consider long paraphrases (where one of the phrases has at least six tokens), and paraphrases that have no syntactic placeholders (a subset of the PPDB contains general syntactic symbols, such as `wishes to be [VP/NP]`). From this set, we randomly sample 4,000 paraphrases. Our sample yields 22,751 aligned same words, 25,973 aligned different words, 2,782 unaligned same words, and 163,474 unaligned different words. We randomly sample 2,500 words from each category. For the aligned words, we only use cases where there is a 1-1 alignment.

To generate word-level embeddings, we run each phrase through a set of transformer models and for each pair of words, we take the cosine similarity between the embeddings of the two words.

3.2.1 Results

Figure 3 shows the distributions of similarity scores for all four sets of words for several models. Same, aligned words consistently have the highest similarity. The other categories tend to overlap. Because we are using paraphrases, we would hope that aligned different words would have higher similarity, but that is not consistently the case.

Comparing the models, there are some notable variations. Between BERT base and BERT large, the biggest shift is that unaligned words that are the same have much lower similarity in BERT large,

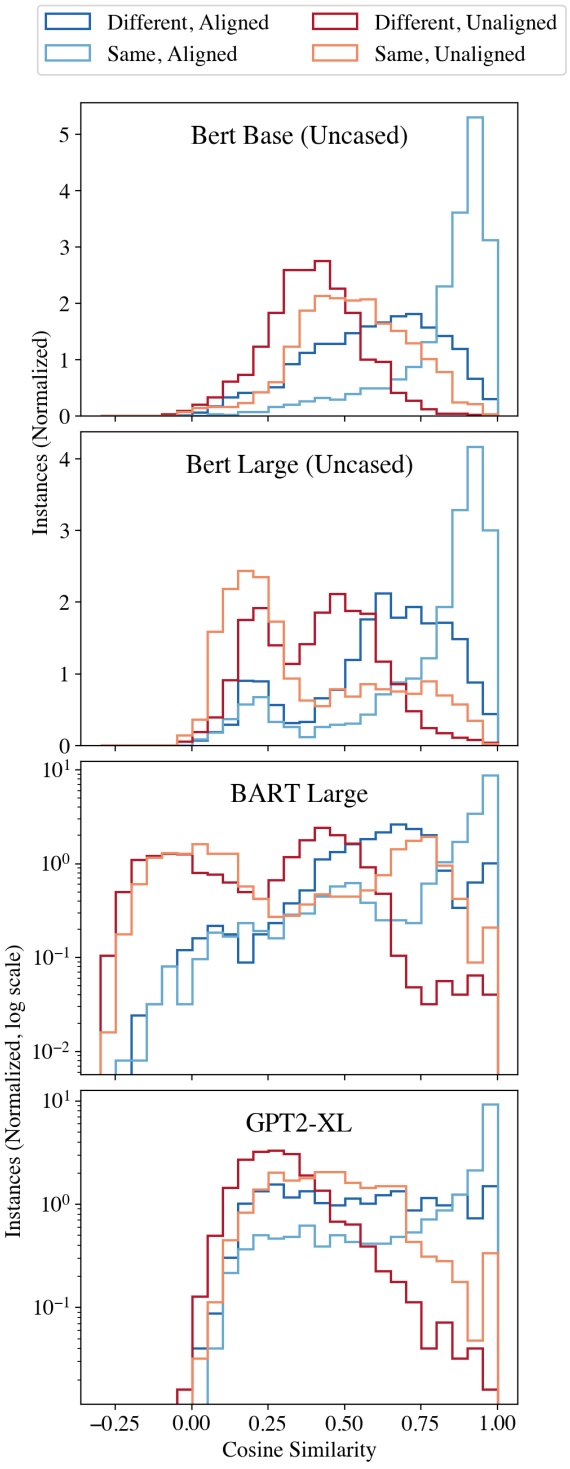

Figure 3: Distributions of cosine similarities for several models for same and different aligned and unaligned words (best seen in color). Cosine similarity is binned into bins of width 0.05. Note that the bottom two graphs use a log scale because the peak at 1.0 makes it hard to see variations otherwise.

though there is also a new peak for aligned words around 0.2. Comparing the two BERT models with BART and GPT-2, there is a much sharper peak for BART and GPT-2 for same aligned words, which is consistent with prior work (Ethayarajh, 2019).

BART is the only model to have a substantial number of negatively correlated word pairs. Many of these involve function words or punctuation. For the unaligned cases, negative cosine similarity is fine because the words should not have the same meaning. For the 26 cases of aligned pairs, it is unclear why the representations are so different. For example, `the` plays the same role in `( , the commission considered` and `, the commission had before it.` Similarly, `aim` and `view` should be very similar in `aim of improving the` and `with a view to improving the.`

Qualitatively looking at examples, we notice that when a token appears in a different position in the paraphrase, the similarity tends to be on the lower end of the distribution (e.g., `action` in the phrases `plans of action for the implementation` and `action plan for the implementation` has a similarity of 0.28). To explore this, we consider 2,181 aligned same words. We measured the cosine similarity of the last layer of BERT broken down by the variation in position (plotted in Figure 7 in the Appendices). Spearman's $\rho = -0.29$ (p-value $< 10e-42$), indicating that similarity decreases for larger changes in position. This supports observations in prior work (Mickus et al., 2020), but now with the knowledge that the overall context has the same meaning. This is not intuitive behavior; because these words are aligned in a paraphrase, we would expect that the position of the word would not substantially affect its representation. This may indicate that the representations are encoding some information about syntactic structure, which can vary without changing semantics.

Conclusions: (1) Contextual word embedding methods consistently handle aligned words in paraphrases, but with substantial variations across models in how peaked the distributions of same-aligned words are. (2) Even when controlling for the meaning of the context, BERT represents words differently depending on their position.

3.2.2 Punctuation

Punctuation is a core part of language that functions quite unlike words; punctuation groups words together or separates them, and contributes to the overall structure and meaning of a phrase or sentence. Punctuation plays an important role in dis-

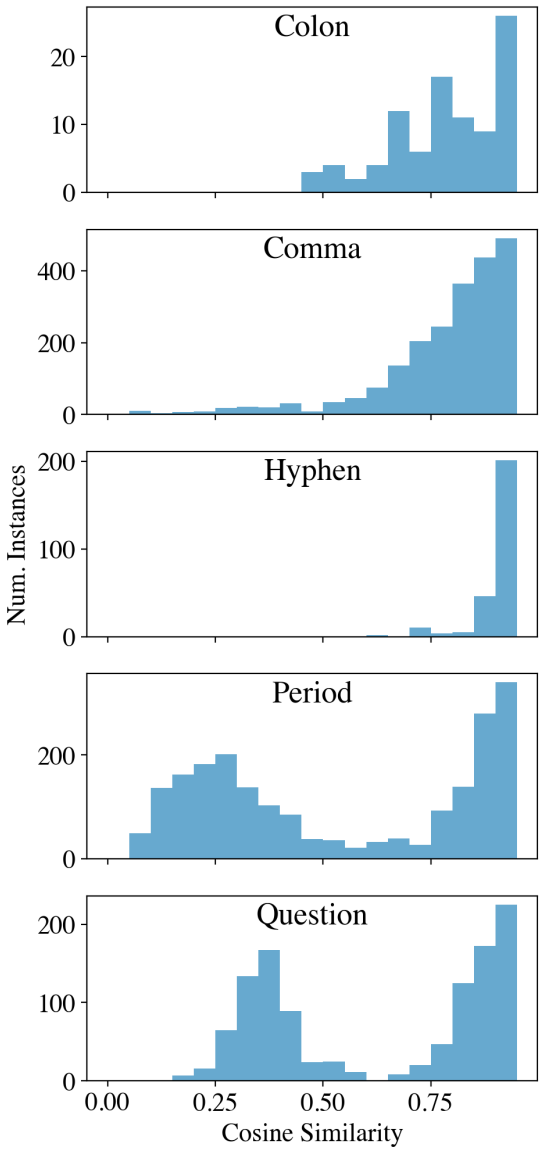

Figure 4: Distribution of cosine similarities using the last layer of BERT for aligned same words, broken down by punctuation mark.

tinguishing between different types of text, such as texts by different authors (Soler-Company and Wanner, 2017) or texts produced by different Twitter communities (Tatman and Paullada, 2017). Embeddings are used to generate punctuation for text that is lacking punctuation, such as recorded transcripts (Yi and Tao, 2019). To explore how BERT handles punctuation, we consider the cosine similarity distribution for different sets of punctuation tokens for the last layer of BERT. We find that punctuation has a broader distribution of cosine similarities than other tokens, indicating that punctuation embeddings vary widely dependent on the surrounding context.

Phrases	Cos. Sim. (Last Layer)
it is important , however , however , it should be	0.08
okay , i 'm sorry oh , i am so sorry	0.84
well , it 's true . this is true .	0.19
, that 's all right . , this is good .	0.15
where have you come from ? where are you from ?	0.94
news - politics - world - news - international politics -	0.71

Table 5: Examples of aligned punctuation marks with varying cosine similarities. The aligned tokens are underlined.

In Figure 4, we break these trends down by individual punctuation marks, focusing only on aligned same words. We look at the most common punctuation marks. Of these punctuation marks, the comma and period show the widest distributions. Even when they play the same role in the paraphrase, they can be given very different embeddings, indicating how highly contextualized these punctuation marks are. The question mark and dash are less contextualized; this is most likely because these punctuation marks are used in more prescribed circumstances. In this dataset, in all but one example, the question mark is the last token; the dash is the first token in all but two examples.

Table 5 shows examples in context of each of these punctuation marks. Looking at the low similarity cases, one common pattern is that the phrase contains a contraction that is expanded in one phrase (e.g., "it is" and "it's").

Conclusion: BERT's representation of punctuation is surprisingly context sensitive, with substantial variation even when we control for meaning.

3.2.3 Polysemy

Previous work has shown that BERT embeddings form clusters based on word senses (Wiedemann et al., 2019). In the context of aligned words in a paraphrase, we would expect even a highly polysemous word to have similar embeddings in the two phrases.

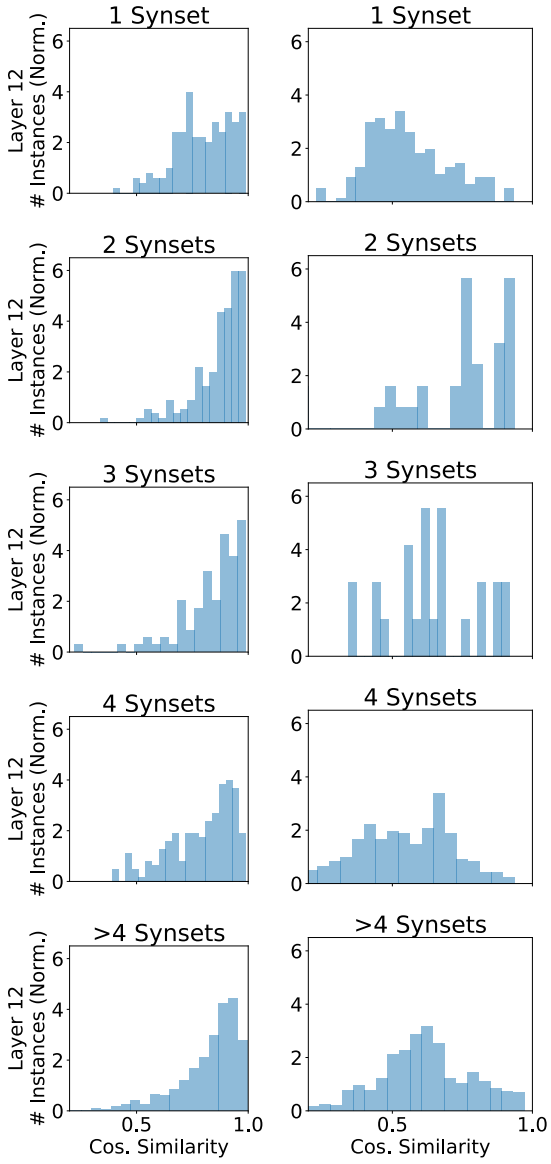

Figure 5: Distributions of cosine similarity for aligned (left) and unaligned (right) same words from the last layer of BERT grouped by the number of senses each word has.

To measure polysemy, we consider the number of WordNet synsets of a word, focusing on same aligned and same unaligned words. In order to have enough data to make a good comparison, we use the 4,000 sampled paraphrases from Section 3.2, as well as an additional random sample of long paraphrases with at least one unaligned same word. We then downsample the aligned same words to get 1,597 instances of both unaligned and aligned same words that are present in WordNet, with up to 52 synsets.[10]

In Figure 5, we show the cosine similarity distributions for both aligned and unaligned words with different levels of polysemy across the last layer of BERT. There is not a substantial difference between words with different synsets, which supports our conclusion that BERT successfully captures the semantics of aligned same words in paraphrases. We do see a difference between aligned and unaligned words. Aligned words peak at a high cosine similarity, while unaligned words roughly follow a normal distribution centered around 0.5. Note that for unaligned words with two or three synsets, there is not enough data to draw conclusions about the cosine similarity distributions. Overall, these plots show that even highly polysemous aligned same words have very similar embeddings in the context of a paraphrase.

Conclusion: How polysemous a word is does not substantially impact BERT's ability to consistently represent it.

[10]We look up WordNet synsets using the Python NLTK library (Bird et al., 2009).

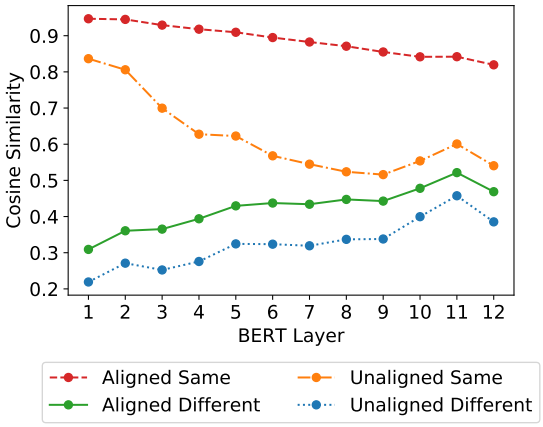

Figure 6: Cosine similarity for different groups of words in the PPDB across all layers of BERT. Decreasing cosine similarity indicates increasing contextualization, and vice versa.

3.3 Contextualization in BERT Layers

In this section, we consider how context-specific the embeddings in a paraphrase are. Ethayarajh (2019) showed that BERT word embeddings are more context-specific in higher layers. They measure this using the self-similarity of words, defined as the average cosine similarity between a word's contextualized representations across its unique contexts, and show that self-similarity consistently decreases with higher layers of BERT, indicating that the contextualization of words is increasing.

We compare this observation to the paraphrase setting that we have been exploring in this paper. Because there are only two phrases in a paraphrase,

we cannot implement the full self-similarity metric. Instead, we measure the cosine similarity between two words aligned in a paraphrase, shown in Figure 6. This revised metric measures is similar to self-similarity.

We see two trends in Figure 6. The first, decreasing cosine similarity, is seen with same words, whether aligned or unaligned, and is similar to what Ethayarajh (2019) report with decreasing self-similarity scores. This trend is stronger with unaligned words than with aligned words, indicating the model is capturing the fact that while these words have the same form, they are being used differently. The second trend that we see is the opposite, increasing cosine similarity, and we see this trend with different words, both aligned and unaligned. This indicates decreasing contextualization.

Conclusion: As seen in prior work, the standard way of using vectors from BERT's layers does not capture the same level of contextualization in all layers. However, in contrast to prior work, when controlling for semantics of the context, it seems that later layers are capturing more of the context, appropriately making words less similar when they are being used in different ways.

4 Conclusion

Paraphrases with word alignments are a useful tool for studying the behavior of contextual language models. In this paper, we used them to study several contextual models, with a particular focus on BERT. Where possible, we compared our results with prior work, finding patterns that are consistent with the literature. Specifically, our results confirm that BERT consistently represents paraphrases, even for cases with polysemous words, but that individual word representations are overly sensitive to position, particularly for punctuation. One exception is that we found that words in a sentence are more similar to each other in later layers of BERT, in contrast to prior work that did not control for meaning using paraphrases.

The analysis method we introduced opens up new opportunities, such as the comparison of aligned and unaligned, same and different words, which shows the sensitivity of these models to the specific word used. Paraphrases have the potential to inform exploration of other representation methods, showing which way of using the output of language models most accurately captures semantics consistently. We hope our findings will inform future work on contextualized models, and the applications that rely on them.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604.

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

Ying-Hong Chan and Yao-Chung Fan. 2019. BERT for question generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 173–177.

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*, pages 4171–4186.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 55–65.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows punta cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss 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)*, pages 3509–3514.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets

of BERT. 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 4365–4374.

Somaia Mahmoud and Marwan Torki. 2020. AlexU-AUX-BERT at SemEval-2020 task 3: Improving BERT contextual similarity using multiple auxiliary contexts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 270–274. International Committee for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition. *Linguistic Data Consortium, Philadelphia*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–566.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992.

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.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Juan Soler-Company and Leo Wanner. 2017. On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 681–687.

Charles Spearman. 1910. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295.

Rachael Tatman and Amandalynne Paullada. 2017. Social identity and punctuation variation in the #bluelivesmatter and #blacklivesmatter twitter communities. In *The 33rd Northwest Linguistics Conference*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 130–139.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Konferenz zur Verarbeitung natürlicher Sprache / Conference on Natural Language Processing*.

J. Yi and J. Tao. 2019. Self-attention based model for punctuation prediction using word and speech embeddings. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7270–7274.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Extra breakdowns of results

Table 6 presents an expanded version of Table 3, with results for each layer of BERT. All layers perform better with longer paraphrases, but the improvement is largest for the last layer

Figure 7 shows the specific values for similarity broken down by distance apart of words in the phrases. This shows the pattern of decreasing similarity as words are further away.

	Average Length			All
Method	1-2.5	2.5-4	4-6	
BERT Layer 1	0.18	0.35	0.47	0.34
BERT Layer 2	0.18	0.35	0.49	0.33
BERT Layer 3	0.18	0.37	0.48	0.31
BERT Layer 4	0.18	0.38	0.48	0.3
BERT Layer 5	0.18	0.39	0.48	0.3
BERT Layer 6	0.19	0.39	0.49	0.29
BERT Layer 7	0.2	0.4	0.49	0.3
BERT Layer 8	0.21	0.4	0.5	0.28
BERT Layer 9	0.21	0.4	0.5	0.28
BERT Layer 10	0.22	0.38	0.48	0.29
BERT Layer 11	0.22	0.36	0.46	0.29
BERT Layer 12	0.1	0.35	0.51	0.16
BERT Concat.	0.2	0.4	0.51	0.31
w2v Average	0.35	0.32	0.41	0.43
PPDB model	0.41	0.50	0.51	0.50
Num. phrases	17,517	5,349	2,870	25,736
Avg. human	2.40	2.94	3.26	2.60

Table 6: This is a version of Table 3 with per-layer results. Spearman's ρ between human-annotated PPDB paraphrases and different embedding methods (BERT and w2v), broken down by average paraphrase length (the average number of words in each of the two phrases in the paraphrase). At the bottom of the table, we include the length distribution of the human-annotated paraphrases, as well as the average human annotation for each set of grouped lengths.

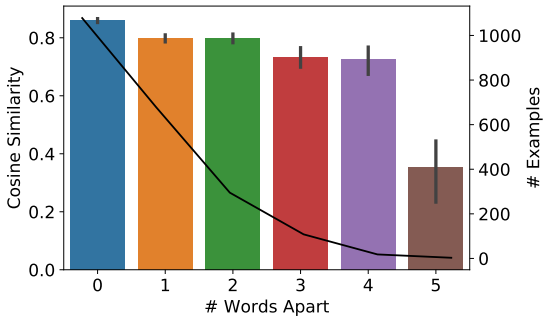

Figure 7: Cosine similarity using the last layer of BERT for aligned same words broken down by the number of words apart the words are in the two phrases (shown in bar plot and left y-axis). Error bars indicate confidence intervals. The line graph and the right y-axis show how many examples we have for each category.