Looking Into the Black Box - How Are Idioms Processed in BERT?

Anonymous ACL submission

Abstract

Idioms such as "call it a day" and "piece of cake" are ubiquitous in natural language. How are idioms processed by language models such as BERT? This study investigates this question with three experiments: (1) an analysis of embedding similarities of idiomatic sentences and their literal spelled-out counterparts, (2) an analysis of word embeddings when the word appears in an idiomatic versus literal context, and (3) an attention analysis of words when they appear in an idiomatic versus literal context. Each of these three experiments analyse results across all layers of BERT. Experiment 1 shows that the cosine similarity of the embeddings of an idiom sentence and its spelled-out counterpart increases the deeper the layer. However, when compared to random controls, layer 8 is where the spelled-out counterpart is ranked highest in embedding similarity. Experiment 2 shows that the embedding of single words in idiomatic versus literal contexts diverge and become the most different in layer 8 also. Experiment 3 shows that other sentence tokens pay less attention to a word inside an idiom compared to the same word in a literal sentence. Overall, the study suggests that BERT "understands" idiomatic expressions, and that it processes them more akin to a syntactic phenomenon than purely a semantic one. A mechanism for this understanding in BERT is attention, which illustrates that idioms are semantically and syntactically idiosyncratic.

1 Introduction

"Why would you put all your eggs in one basket? I can't wrap my head around it." - idioms such as "put all one's eggs in one basket" and "wrap one's head around" are used frequently in natural conversations. Despite their abundance, much remains to be explored regarding their syntactic, semantic, and pragmatic characteristics, and how they are processed by the human brain as well as NLP models. Recent Transformer-based language models such as BERT have demonstrated strong capabilities in a sweep of tasks involving natural language understanding. However, few attempts have been made to understand the inner workings of BERT in terms of idiom processing. In this study, we conduct three experiments to explore how BERT processes idiomatic sentences - we explore embeddings on the sentence level and on the word level, with and without context; we also explore the attention from other sentence tokens to a word inside an idiom compared to a literal context. The results shed light on how BERT processes idioms.

1.1 Research Questions

In this study we explore three questions:

- How does BERT represent idiomatic sentences as opposed to their literal spelled-out counterparts across different layers in the network? For example, "Birds of a feather flock together" versus "People with similar interests stick together".
- How does BERT represent a *word* inside an idiom compared to the same word in a literal context? For example, the word "feather" in "Birds of a feather flock together" versus "My parakeet dropped a green feather."
- What is the mechanism by which the network processes idioms?

We hypothesise that if BERT "understands" idioms, sentence embeddings of idiom sentences and their literal spelled-out counterparts would become more similar across layers. We also hypothesise that, if idiomatic expression is purely a semantic phenomenon, and if semantic representation is concentrated in the upmost layers, word embeddings of a word inside an idiom and the same word in a literal context would diverge the deeper the layer, and become the most different in the upmost layers. In terms of mechanism, because idioms often act as single units, we hypothesise that a word inside idioms would receive less attention from the rest of the sentence compared to the same word in a literal sentence.

1.2 Related Work

The current study is related to linguistic research on idioms, research on the inner workings of BERT, often coined "BERTology", and more specifically BERT's processing of idiomatic expressions.

Linguistics of idioms: Idioms seem easy to spot but difficult to define. They are conventionalised, affective, inflexible, and often figurative multi-word expressions used primarily in informal speech. Some theories suggest that that idioms are lexically, syntactically, semantically and pragmatically idiosyncratic (Baldwin and Kim, 2010). Syntactically, idioms can function as noun phrases, verb phrases or clauses. Semantically, an idiom has a phrasal entry in the lexicon, associated directly with a single semantic representation. Idioms are often non-compositional - the meaning of an idiom often cannot be predicted based on the meaning of the words it is composed of (Nunberg et al., 1994).

BERT and BERTology: BERT (Devlin et al., 2018) is a large Transformer network pre-trained on 3.3 billion tokens of written corpora including the BookCorpus and the English Wikipedia (Vaswani et al., 2017). Each layer contains multiple self-attention heads that compute attention weights between all pairs of tokens in the input. Attention weights can be seen as deciding how relevant every token is in relation to every other token for producing the representation on the following layer.

In terms of how language structure is represented in BERT, Jawahar et al. (2019) observed that different layers encode different linguistic information. Lower layers capture phrase-level information (i.e. surface features), middle layers capture syntactic information and higher layers capture semantic features.

Studies disagree on where and how much semantic information is encoded. For example, Tenney et al. (2019) suggest that semantics is spread across the entire model. Mickus et al. (2020) suggests that BERT capture semantic similarity between words better than sentence-level coherence. Lenci et al. (2021) explored word-level semantic representation in BERT as well, but for out-of-context words. It was found that the uppermost layer (the most contextualised layer) was in fact the worst-performing, globally.

Idiom processing in BERT: The processing of idiomatic expressions in BERT is under-explored so far and is considered a challenge (Salton et al., 2014). Nedumpozhimana and Kelleher (2021) investigated how BERT recognises idiomatic expressions in a sentence using a masking task. They suggested that BERT's idiomatic expression indicator is found both within the expression itself and in the surrounding context. Moreover, BERT can detect semantic disruption in a sentence caused by idiomatic expressions. However, this study focused on analysing and aggregating embeddings in the final layer only, and did not investigate how representations change across different layers.

2 Experiments

To look into the black box of how BERT processes idiomatic language, we conducted three experiments to assess sentence embeddings, word embeddings and attention across all layers of the network.

2.1 Experiment 1: Idiom versus Spelled-out sentence embedding analysis

Experiment 1 investigates how sentence embeddings of idiomatic sentences evolves across layers.

Dataset: We manually curated 100 idioms in English. For each idiom, we created an idiom sentence, as well as a spelled-out counterpart, which expresses the meaning of the idiom sentence in literal language. For example:

- **Idiom :** one's two cents
- **Idiom sentence :** You can put in your two cents later.
- **Spelled-out sentence:** You can share your thoughts later.

2.1.1 Methods and Results

To embed the sentences, we used the library Transformers from Huggingface (Wolf et al., 2020) and the medium-sized BERT model (`bert-base-cased`) which contains 12 layers, 12 attention heads, and a total of 110M parameters. Let $\mathcal{S}$ denote the dataset of all (idiom, and spelled-out) sentence tuples (in the notations below we represent idiom sentences with s_i, and spelled-out sentences with s_s).

We determine whether BERT's representation of an idiom sentence is similar to its spelled-out counterpart using two metrics:

- Metric 1: the *raw cosine similarity* $\phi(s_i, s_s) = \frac{s_i \cdot s_s}{\max(||s_i||_2 \cdot ||s_s||_2, \epsilon)}$ computed for all $(s_i, s_s) \in \mathcal{S}$.
- Metric 2: the *cosine similarity ranking* computed for all (s_i, s_s) with $(s_i, s_s) \in \mathcal{S} \times \mathcal{S}$.

The raw cosine similarity in Metric 1 indicates the how close an idiom and spelled-out pair is in the embedding space, while the similarity *ranking* in Metric 2 determines the quality of an embedding in capturing semantic nuances compared to controls. A close idiom and spelled-out pair relative to controls should converge to a high rank. The reasoning is that when an idiomatic sentence s_i is compared against all spelled-out sentences s_s in the dataset, its spelled-out counterpart should be the most similar in semantic content. If its similarity rank is high, it means that the embeddings encode the semantic information that allows the ranking to *disambiguate* the correct spelled-out counterpart among all sentences.

Cosine Similarity: We aggregate the activations of all sentence tokens into a single flattened vector[1]. We calculate the cosine similarity between each idiom sentence and its spelled-out counterpart. As a baseline, we calculate the cosine similarity between an idiom sentence and a random spelled-out sentence. In all cases, we report the mean cosine similarity.

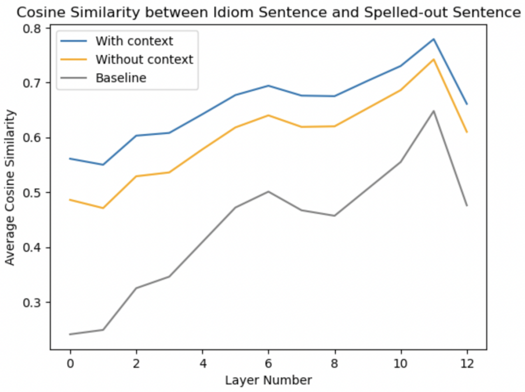

Figure 1: Experiment 1 - Cosine similarity of Idiom and Spelled-out sentence pairs

The results are shown in Figure 1. Overall, the cosine similarity between idiom sentence and its spelled-out counterpart is higher than the random baseline. Cosine similarities between an idiom sentence and its spelled-out counterpart changes from on average 0.56 to 0.78 from layer 1 to layer 11 (two identical sentences have a cosine similarity of 1). All sentence embeddings first drop on layer 1, then become more similar across layers, peaking in layer 11. Similar patterns were reported by Wang and Kuo (2020) and Tian et al. (2021). However, we cannot conclude that layer 11 is where BERT recognises the idiomatic and literal sentences to be the most similar, due to the fact that the cosine similarity to random controls in baseline also peaks at layer 11 (grey line). For this reason, we employed a similarity *ranking* metric to further evaluate our hypothesis.

Idiom and Spelled-out sentence pair ranking: In order to determine how similar a spelled-out counterpart is to its idiom sentence compared to controls, we computed the *rank* of the spelled-out counterpart among 100 sentences in cosine similarity.

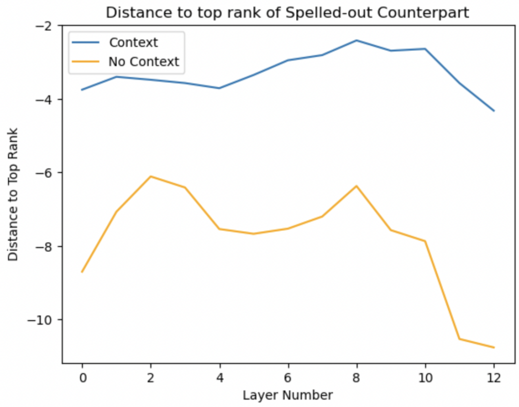

Figure 2: Experiment 1 - Similarity ranking, where we plot the similarity *ranking* of the spelled-out counterpart - the closer to zero, the more similar the spelled-out counterpart is to the idiom sentence compared to controls.

The pair ranking results can be observed in the blue line in Figure 2. The graph shows that the ranking of the spelled-out counterpart is relatively high from early layers: average rank 4 (out of 100) on layer 1, peaking at rank 2 (out of 100) at layer 8. This suggests that BERT recognises the surface form of idioms and integrates them early on in the network. On the other hand, layer 11 ranks lower in similarity than some of the earliest layers. This

[1] In order to calculate the cosine similarity between two sentences of different lengths, we pad the shorter sentence in each pair with [PAD] so that the two have the same number of tokens.

suggests that idioms are processed and integrated by middle layers of BERT rather than in the final layers which are usually associated with semantic representation. However, as the idiom and literal sentences share some of the context text, the high ranking might have been artificially boosted by text overlap. In order to remove this confound, we edited the dataset by removing shared context and ran a followup to Experiment 1.

2.2 Removing Context

Conscious that the surrounding language in our idiom sentences might be influencing the results, we conducted a follow-up experiment in which supplementary or contextual elements were removed. 48 out of 100 of our sentences were adapted in this way to reduce this influence. For instance,“I’m tired, why don’t we call it a day” was changed to “Why don’t we call it a day” in this followup study. We then repeated the above process with the same two metrics: cosine similarity and pair ranking.

We followed the same methods as the original Experiment 1, and found the same pattern when context was removed (Figure 1). The average cosine similarity is slightly lower than that of the original data, but the pattern across layers remains the same.

Pair *ranking* of context-removed data yielded similar results to the pair ranking of the original data, shown in Figure 2. In this case, the similarity ranking starts at average rank 9 (out of 100) at layer 0, peaks on layer 2 and layer 8 with rank 6 (out of 100), then declines from layer 9. The original and followup experiments suggest that idioms are best “understood” by the middle layers of BERT that are usually associated with syntactic processing.

2.3 Experiment 2: How does the embedding of a word within an idiom change compared to the same word in a literal context

In Experiment 1, we saw that *sentence* embeddings capture idioms by the middle layers of BERT. Experiment 2 investigates how *word* embeddings change when the word is in an idiomatic versus literal context.

Dataset: For each Idiom sentence we manually created an unrelated literal sentence that contains a word from the associated idiom. For example:

- *Idiom sentence*: Don’t beat around the [bush].
- *Unrelated literal sentence*: There’s a small [bush] in the garden.
- *Word*: bush

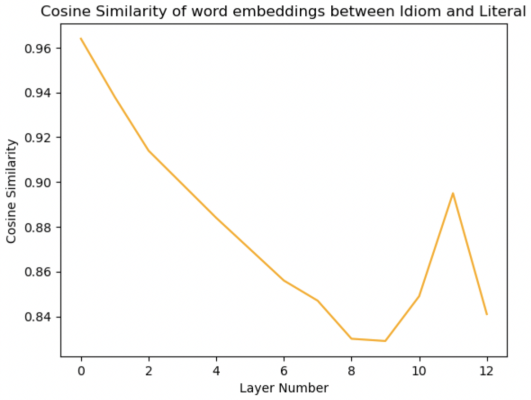

Figure 3: Experiment 2 - Cosine similarities of word embeddings between idiomatic and literal use of the word

Methods and Results: We identified the index of the target word after the sentences were tokenised, and retrieved the embedding for this word across all layers of BERT.

Figure 3 depicts the cosine similarity between the embedding of the word in the Idiom sentence versus its unrelated literal control. The results show that the embedding of the target word (e.g. “bush”) between idiom and literal contexts are identical in layer 0 (because they are the same token). They then diverge steadily across the layers, and become the most *dissimilar* in layers 8 and 9, before rising again from layer 10 and dropping in layer 12. This shows that BERT represents the target words most differently in its mid to late layers, as opposed to its uppermost layers, echoing the findings of experiment 1.

2.4 Experiment 3: Does BERT pay different attentions to words inside idioms versus literal context

Experiment 1 and 2 show that sentence embeddings of idiom sentences become the most similar to their spelled-out counterparts in middle layers, and word embeddings of a word between an idiomatic context and a literal context become the most *dissimilar* in the mid to late layers. These results suggest that BERT treats the words in idioms differently compared to words in a literal context. What is the mechanism that allows the network to “understand” idioms? As self-attention is central to the power of Transformer models, we hypothesise that the network integrates idioms by paying different attention when a word is in an idiom versus a literal context. Specifically, we hypothesise that words inside idioms are less connected to the rest of the sentence because the whole expression functions as a single unit.

2.4.1 Methods and Results

Experiment 3 compares the attention to a word inside an idiom with attention to the same word in a literal context. For each idiom sentence, we select a word inside the idiom, and create an literal control sentence that is unrelated in meaning. For example:

- *Idiom sentence*: Why don't we call it a [day].
- *Literal sentence*: I will arrive the [day] after tomorrow.
- *Target word*: day

We identified the indices of the target word (e.g. "day") in the idiom and the literal sentence. Then for each sentence and for each layer, we calculated the average attention from all other sentence tokens to the target word.

Figure 4 plots the average attention in each layer of BERT. Overall, we see that a sentence pays *less* attention to a word inside an idiom than it does to the same word in a literal context. The difference is most significant in layer 8, where attention to the target word is the lowest for idiom sentences.

Experiment 3 provides further evidence that BERT "understands idioms" - it pays different attention to words inside an idiom compared to when those words are in a literal context. The difference is the biggest in *layer 8*, repeating the pattern of Experiment 1 and 2. The results support the idea that idioms are less compositional, and BERT integrates them into sentences as idiosyncratic units.

3 Future Studies

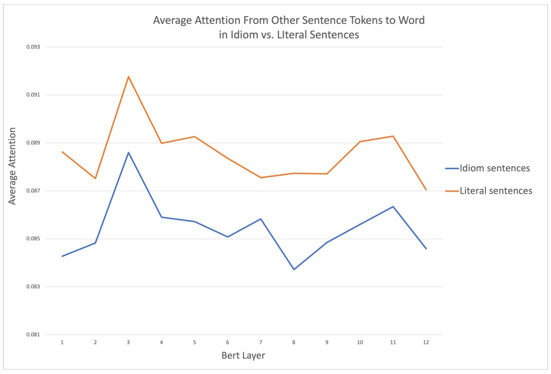

Figure 4: Experiment 3 - Attention from other sentence tokens to word inside an idiom sentence versus a literal sentence

Linguistics research debates on whether all idioms are non-compositional, and further research could test whether this holds true for BERT. The "idiom decomposition hypothesis" (Gibbs et al., 1989) suggests that idioms being decomposable or non-decomposable is significant to how they are processed. An idiom is decomposable if its meaning can be deduced from the individual words that form it. Our dataset contains examples of both decomposable and non-decomposable idioms, with an unbalanced weighting towards the former. An example of a decomposable idiom is: "fat chance of that happening", whereas an example of a non-decomposable idiom is: "kick the bucket". A future study looking at whether decomposability affects how BERT processes idioms, and in which layers this can be most observed, would build on the current study's findings.

Another area of future study is comparing the processing of different types of idioms. It was mentioned earlier that the results of Experiment 3 suggest that BERT pays different attention to words in idioms compared to words used in their literal context. We could further assess BERT's tendency to pay different levels of attention to different degrees of literalness by comparing various types of idioms. Idioms vary in their semantic opacity, which affects the rigidity of their composition (O'grady, 1998). Words in highly opaque idioms (e.g. "pull strings") tend to be metaphorical and thus are often irreplaceable. On the other hand, less opaque idioms (e.g. "should have one's head examine") allow variability in lexical items, such as substitution of "should" with "need to". Idioms vary syntactically as well, facing different constraints depending on whether they are verbal (e.g. "kick the bucket"), nominal (e.g. "tooth and nail"), or sentential (e.g. "the fat is in the fire") (O'grady, 1998), such as whether insertion of quantifiers is permitted or not. As our experiment results suggest that idioms are also processed syntactically in BERT, it would also be interesting to conduct a further study with higher coverage of different syntactic types of idioms (Tan and Jiang, 2021).

4 Discussion

We investigated how BERT processes idioms across its layers on a sentence level and word level. Experiment 1 shows that on a sentence level, BERT represents an idiom sentence to be more similar to its literal spelled-out counterpart, and this similarity peaks in layer 8. A similar pattern was found when context was removed. Experiment 2 shows that on a word level, BERT represents a word inside an idiomatic versus a literal context increasingly differently across layers, peaking in layers 8 and 9. Experiment 3 shows that BERT pays different attention to words in an idiom compared to a literal context - words in an idiom receive *less* attention from the rest of the sentence and thus have a weaker link to words outside of the idiom.

Overall, our experiments have demonstrated that BERT is capable of “understanding” idioms with and without surrounding context, which is in line with findings from Nedumpozhimana and Kelleher (2021). Returning to the question of whether BERT processes idioms as a purely semantic or syntactic phenomenon: previous findings (e.g. Jawahar et al. (2019) and Mickus et al. (2020)) suggest that semantic information is primarily handled by the deepest layers of BERT, and the last layer (12) is the most frequently used embedding layer for NLU tasks. In comparison, middle layers are associated with syntactic processing. For example, Jawahar et al. (2019) found that layer 8 is best for tasks such as subject-verb agreement and auxiliary classification. In this context, we suggest that idioms are processed *not* as a purely semantic phenomenon but rather more akin to other syntactic features. This is likely due to the fact that words in idioms not only bring different meaning, but are also integrated with the rest of the sentence differently - they stick together as a single unit and share a weaker syntactic link with words outside the idiom.

5 Conclusion

Idiomatic expressions are part and parcel of everyday language use. This study shows that BERT is capable of understanding idiomatic expressions with and without surrounding context. The processing is handled more akin to a syntactic feature than a purely semantic one. The results of this study raise the questions of which characteristics of idioms are considered semantic variations by BERT, and whether the last layer of BERT is always the most effective at capturing semantic meaning.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *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)*.

Raymond Gibbs, Nandini Nayak, and Cooper Cutting. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, 28(5):576–593.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2021. A comprehensive comparative evaluation and analysis of distributional semantic models. *arXiv preprint arXiv:2105.09825v1*.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, bert? assessing bert as a distributional semantics model. *Proceedings of the Society for Computation in Linguistics*, 3(34).

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT’s idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

William O’grady. 1998. The syntax of idioms. *Natural Language Linguistic Theory*, 16:279–312.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

Ye Tian, Tim Nieradzik, Sepehr Jalali, and Da-shan Shiu. 2021. How does bert process disfluency? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 208–217.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Bin Wang and C-C Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 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.