How Are Idioms Processed Inside Transformer Language Models?

Anonymous ACL submission

Abstract

Idioms such as "call it a day" and "piece of cake", are ubiquitous in natural language. How do Transformer language models process idioms? This study examines this question by analysing three models - BERT, Multilingual BERT, and DistilBERT. We compare the embeddings of idiomatic and literal expressions across all layers of the networks at both the sentence and word levels. Additionally, we investigate the attention directed from other sentence tokens towards a word within an idiom as opposed to in a literal context. Results indicate that while the three models exhibit slightly different internal mechanisms, they all represent idioms distinctively compared to literal language, with attention playing a critical role. These findings suggest that idioms are semantically and syntactically idiosyncratic, not only for humans but also for language models.

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 large language models have demonstrated strong capabilities in a sweep of tasks involving natural language understanding (e.g. Brown et al. (2020)). However, few attempts have been made to understand the inner workings of these language models in terms of idiom processing. In this study, we conduct three experiments to explore the inner workings of transformer language models in idiom processing. Specifically, we investigate the processing of BERT, Multilingual BERT and DistilBERT by comparing the embeddings on the sentence level and on the word level. We also explore the attention mechanism on idioms compared to literal contexts. We ask three questions:

- How do Transformer language models (LMs) represent idiomatic sentences as opposed to their literal spelt-out counterparts across different layers in the network? For example, "Birds of a feather flock together" versus "People with similar interests stick together".
- How do LMs 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."
- How do LMs pay attention to a word inside an idiom compared to a literal context?

1.1 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.

Linguistic theories of idioms: Idioms seem easy to spot but difficult to define. They are conventionalised, affective, and often figurative multi-word expressions used primarily in informal speech (Baldwin and Kim, 2010). Idioms are non-compositional - their meanings often cannot be predicted based on the words they is composed of (Nunberg et al., 1994). Sinclair and Sinclair (1991) postulate that humans process idioms by treating them as a "single independent token".

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. Attention weights can

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179
July 13-14, 2023 ©2023 Association for Computational Linguistics

be seen as deciding how relevant every token is in relation to every other token for producing the representation on the following layer (Clark et al., 2019).

Many studies have explored how different linguistic information is represented in BERT (Mickus et al., 2020; Jawahar et al., 2019; Tenney et al., 2019). 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) suggests that semantics is spread across the entire model. Lenci et al. (2021) found that the uppermost layer in BERT was the worst-performing in downstream tasks. So far, there has been less research on the inner workings of DistilBERT (Sanh et al., 2019) and Multilingual BERT (Pires et al., 2019). Most studies focus on comparing performance cross-lingually or in downstream tasks between these models (Ulčar and Robnik-Sikonja, 2021; Wu and Dredze, 2020; Sajjad et al., 2021; Lenci et al., 2021).

Idiom processing in Language Models: Studies are becoming increasingly engaged with the challenge of idiom representation in language models (Socolof et al., 2021; Garcia et al., 2021b; Dankers et al., 2022). Nedumpozhimana and Kelleher (2021) investigated how BERT recognises idioms, suggesting that the indicator is found both within the expression and in the surrounding context. Madabushi et al. (2021) explored how various input features (e.g. the effect of different problem setups - zero-shot, one-shot, and few-shot) affect LMs' ability to represent idioms. Both studies analyse the aggregated embeddings in the final layer, and do not investigate how representations vary across different layers. Garcia et al. (2021a) probed the representation of noun compounds in LMs, varying in compositionality, in order to assess the retention of idiomatic meaning. Our paper follows a similar paradigm but includes an attention analysis. Finally, Dankers et al. (2022) analysed idiom processing for pre-trained neural machine translation Transformer models from English to seven European languages and found that when the model produces a non-literal (intended) translation of the idiom, the encoder processes idioms more as single lexical units compared to literal expressions.

2 Experiments

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

2.1 Dataset

We utilised the idioms from the EPIE dataset (Saxena and Paul, 2020) to obtain a list of 838 English idioms that occur frequently in language. We then created sentences for the following conditions: for each idiom, we created (1) a sentence containing that idiom, (2) a spelt-out sentence expressing the same idiom in literal language, and (3) two unrelated literal sentences containing a key-word from the idiom (for experiment 2). An example of a datapoint[1]:

- **Idiom :** under the weather
- **Idiom sentence :** I'm feeling under the weather today.
- **Spelt-out meaning:** I'm feeling unwell today.
- **Unrelated literal sentence 1:** Today's weather is nice.
- **Unrelated literal sentence 2:** The weather is meant to change at 10am today.

2.2 Experiment 1: Idiom versus Spelt-out sentence embedding analysis

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

2.2.1 Methods and Results

To embed the sentences, we used the Python library Transformers from Hugging Face (Wolf et al., 2020). We used the medium-sized BERT model (`BERT-base-uncased`), Multilingual BERT (`BERT-base-multilingual-uncased`), and DistilBERT (`distilBERT-base-uncased`). The first two models contain 12 layers and 12 attention heads, while DistilBERT contains 6 layers and 12 attention heads. Let S denote the dataset of all (idiom, and spelt-out) sentence tuples (in the notations below we represent idiom sentences with s_i, and spelt-out sentences with s_s).

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

[1]The entire dataset is released with the paper.

- 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 how close an idiom and spelt-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 (all other non-counterpart spelt-out sentences). A close idiom and spelt-out pair relative to controls should converge to a rank close to 0. The reasoning is that when an idiomatic sentence s_i is compared against all spelt-out sentences s_s in the dataset, its spelt-out counterpart should be the most similar in semantic content.

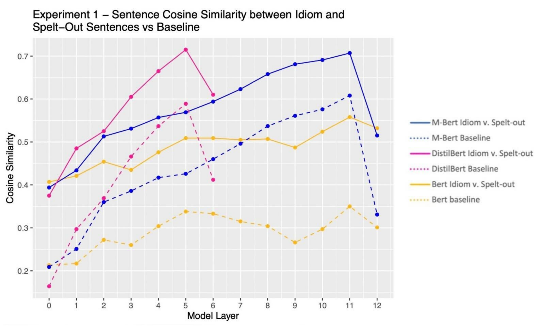

Figure 1: Experiment 1 - Sentence Cosine similarity of Idiom and Spelt-out sentence pairs

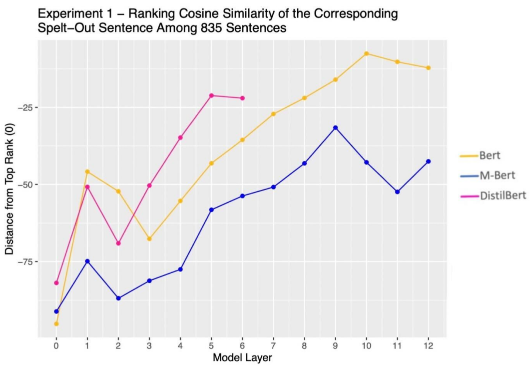

Figure 2: Experiment 1 - Similarity ranking, where we plot the similarity *ranking* of the spelt-out counterpart - the closer to zero, the more similar the spelt-out counterpart is to the idiom sentence compared to controls.

The results are shown in Figure 1 and Figure 2. Overall, the cosine similarity[2] between idiom sentence and its spelt-out counterpart is higher than the random baseline for all three models. For all three models and for every layer in each model, there was a significant difference (all p-values < 0.001) in sentence cosine similarity. Moreover, the t-values increased in deeper layers, which shows that these layers better processed semantic similarities between idioms and their spelt-out counterparts, supporting our hypothesis that the semantic meaning of idioms is captured in deeper layers of BERT.

Among the three LMs, the patterns of DistilBERT and Multilingual BERT most resemble each other, with similarity rising steadily, peaking on the penultimate layer, and dropping on the last layer. In order to evaluate if the LMs represent a literal spelt-out sentence to be *more* similar to random controls, we evaluated a similarity *ranking* metric.

The pair ranking results (Figure 2) show that similarity ranking reaches the highest point in mid to late layers for all 3 LMs, peaking at layer 10 for BERT, at layer 9 for Multilingual BERT and at layer 5 (penultimate layer) for DistilBERT.

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

Experiment 2 investigates how *word* embeddings change for words in idiomatic versus literal contexts. To do this, we see how the the cosine similarity of the embedding of a word inside an idiom versus in a literal context changes across layers, and compare that with a baseline cosine similarity where the word appears in two literal contexts.

Dataset: For each idiom sentence we manually created two unrelated literal sentences which contain a word from the associated idiom. For example, idiom sentence: *Don't beat around the [bush]*. Unrelated literal sentences: (1) *There's a small [bush] in the garden*, and (2) *The dog jumped over the [bush]*. Target word: "*bush*".

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

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

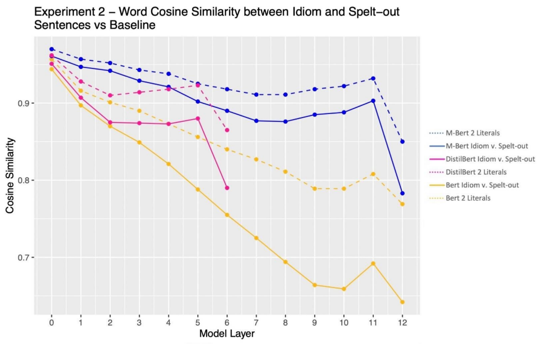

Figure 3: Experiment 2 - Cosine similarities of word embeddings between idiomatic and literal uses of the word

all layers for the idiom sentence and the two unrelated literal sentences. We calculated the cosine similarity for the word embedding (1) between idiom and literal contexts and (2) between the two literal contexts as a baseline.

Figure 3 shows that for all three language models, the similarity of word in two literal contexts (dotted line) is higher than that between idiom and literal contexts (solid line). Like in experiment 1, DistilBERT and Multilingual BERT resemble each other in their patterns. For BERT, the similarity of word embedding between literal and idiom contexts drops significantly more than between two literal contexts. T-test results showed the same pattern observed in experiment 1 as well; there was a significant difference (all p-values < 0.001) in cosine similarity in every layer for all three models, and the absolute value of t-value increased in deeper layers. This confirms our hypothesis that the semantic meaning of idioms is captured in deeper layers of BERT, where words inside idiom drift further from their literal meaning. We see a similar but reduced pattern in Multilingual BERT and DistilBERT.

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

Experiment 1 and 2 show that LMs treat idioms differently to literal expressions. What is the mechanism that allows the networks to process this difference? 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, following the linguistic theory that idiomatic expressions function as a single unit (Sinclair and Sinclair, 1991).

2.4.1 Methods and Results

For each idiom sentence, we selected a word inside the idiom and the indices of the target word (e.g. "bush") in both 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.

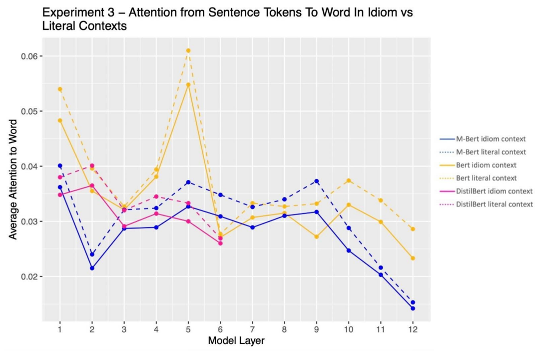

Figure 4: Experiment 3 - Attention from other sentence tokens to word inside an idiom sentence versus a literal sentence

Figure 4 plots the attention in each layer of LMs from all other sentence tokens to the target word. For all three language models, sentence tokens pay *less* attention to a word inside an idiom (solid lines) than they do to the same word in a literal context (dotted lines), meaning that words inside idioms interact less with the rest of the sentence compared to words in literal contexts. Like in experiment 1 and experiment 2, there was a significant difference between attention to a word inside an idiom and that inside a literal context in each layer in all three models (p-values < 0.01). This supports the idea that LMs see idioms as more idiosyncratic units. However, while DistilBERT and Multilingual BERT showed a similar trend in t-values that decreased in degree in the last 2 layers, BERT did not show any particular pattern in t-statistics. Once again we observe that DistilBERT and Multilingual BERT share a similar pattern, whereas BERT displays more variations across its layers.

3 Results Summary

We investigated how Transformer LMs process idioms across their layers on a sentence level and a word level. Experiment 1 shows that on a sentence level, LMs represent an idiom sentence to be similar to its literal spelt-out counterpart. Experiment 2 shows that on a word level, LMs represent a word inside an idiom versus a literal context differently across layers. Experiment 3 shows that 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, echoing the findings of Dankers et al. (2022). All of these results hold across BERT, Multilingual BERT and DistilBERT. We also observe slight differences between the three LMs, with DistilBERT and Multilingual BERT resembling each other in their internal workings more closely than they each do with BERT. In future work we will investigate this phenomenon in models with different architectures, for example GPT and XLNet.

4 Conclusion

Idiomatic expressions are part and parcel of everyday language use. This study investigates the inner workings of idiom processing in three Transformer language models. Results show that LMs represent idioms differently to literal language. Words inside idioms receive less attention compared to words in literal contexts, supporting the linguistic theory that idioms are idiosyncratic even for language models.

A Limitations

While this work sheds light on how language models process idioms, we recognise that experimentation at present has been constrained to BERT. As mentioned in section 3, we aim to probe our findings further by repeating these experiments on a wider range of model architectures, such as GPT, Flan-T5, and LLaMA. Additionally, we recognise that our current dataset only contains English idioms; it would be interesting to extend this to include other languages for future studies.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *?*

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. Association for Computational Linguistics.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

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)*.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

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*.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *CoRR*, abs/2109.04413.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2021. On the effect of dropping layers of pre-trained transformer models. *Journal of Computer Speech and Language*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions.

John Sinclair and Les Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press, USA.

Michaela Socolof, Jackie Chi Kit Cheung, Michael Wagner, and Timothy J O'Donnell. 2021. Characterizing idioms: Conventionality and contingency. *arXiv preprint arXiv:2104.08664*.

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

Matej Ulčar and Marko Robnik-Sikonja. 2021. Training dataset and dictionary sizes matter in bert models: the case of baltic languages.

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*.

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.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? pages 120–130.