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 are idioms processed by Transformer language models? This study investigates this question on three models - BERT, Multilingual BERT and DistilBERT. We compare the embeddings of idiom and literal expressions across all layers of the networks on the sentence level and on the word level. We also explore the attention from other sentence tokens towards a word inside an idiom compared to a literal context. Results show that the three models have different inner workings, but they all represent idioms differently to literal language, with attention being a crucial mechanism. The 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 language models such as BERT have demonstrated strong capabilities in a sweep of tasks involving natural language understanding. (Ref??) 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, M-BERT (Multilingual BERT) and DistilBERT by comparing the embeddings on the sentence level and on the word level. We also explore the attention from other sentence tokens to a word inside an idiom compared to a literal context. 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 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). Sinclair and Sinclair (1991) postulates 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 be seen as deciding how relevant every token is in relation to every other token for producing the representation on the following layer.

Many studies have explored how different linguistic information is represented in BERT (Mickus et al., 2020; ?; 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) suggest that semantics is spread across the entire model. Lenci et al. (2021) found that the uppermost layer in BERT was the worst-performing, globally. There is less work on the inner workings of DistilBert(Sanh et al., 2019) and M-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 BERT: The processing of idiomatic expressions in BERT is under-explored and is considered a challenge (Salton et al., 2014). Nedumpozhimana and Kelleher (2021) investigated how BERT recognises idiomatic expressions, suggesting that the idiomatic expression indicator is found both within the expression and in the surrounding context. This study analysed the aggregated embeddings in the final layer, and did not investigate how representations change across different layers.

2 Experiments

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

2.1 Dataset

Two annotators (native speakers of English) researched the most frequently used idioms in the English language, and manually constructed a dataset of 200 unique idioms[1]. We chose to limit our dataset to 200 idioms to ensure that the idiomatic expressions we test are not too obscure. We did not include idioms wherein the keyword does not have a common literal usage. For instance, we did not include the idiom "in a nutshell", as the word "nutshell" is not frequently used outside of its idiomatic context. Each idiom comes with (1) a sentence containing that idiom, (2) a spelt-out sentence expressing the same in literal language, and (3) two unrelated literal sentences containing a key-word from the idiom. We release our dataset as one of the contributions of this paper. An example of a datapoint:

- **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 evolves across layers.

2.2.1 Methods and Results

To embed the sentences, we used the python library Transformers from Huggingface (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`), all of which contains 12 layers, 12 attention heads.Let $\mathcal{S}$ 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 BERT's representation of an idiom sentence is similar to its spelt-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 $(s_i, s_s) \in \mathcal{S} \times \mathcal{S}$.

[1]To our knowledge, a comparable dataset with these features does not exist. While recent work is beginning to address the scarcity of multiword expression datasets, for instance the EPIE dataset which contains formal and static idioms (Saxena and Paul, 2020), an idiom-focused dataset that allows for both sentence-level and word-level analysis is lacking.

The raw cosine similarity in Metric 1 indicates the 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. A close idiom and spelt-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 spelt-out sentences s_s in the dataset, its spelt-out counterpart should be the most similar in semantic content.

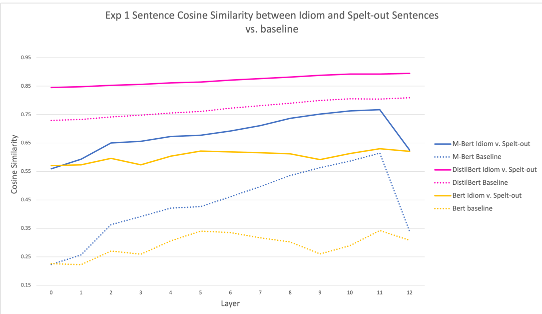

Figure 1: Experiment 1 - Sentence Cosine similarity of Idiom and Spelt-out sentence pairs

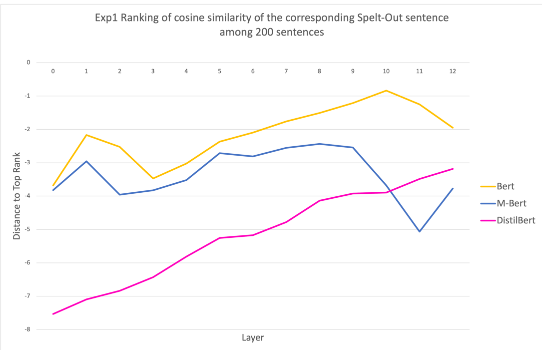

Figure 2: Experiment 1 - Similarity ranking, where we plot the similarity *ranking* of the spelt-out counterpart - the closer to zero, the more similar the spelt-out counterpart is to the idiom sentence compared to controls.

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. Interestingly, DistilBert has much higher raw sentence similarity for both idiom-literal pairs *and* for random baselines; it also has less variation across layers compared to the other two models. 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 increases across layers, peaking at layer 10 for BERT, at layer 8 for Multilingual Bert and at layer 12 for DistilBert. BERT performs the best and DistilBert the worst. Once again we observe significant differences for 3 models. Overall, experiment 1 show that LMs are able to discern idiom expressions on a sentence level.

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 when the word is in an idiomatic versus literal context.

Dataset: For each idiom sentence we manually created two unrelated literal sentences that 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*".

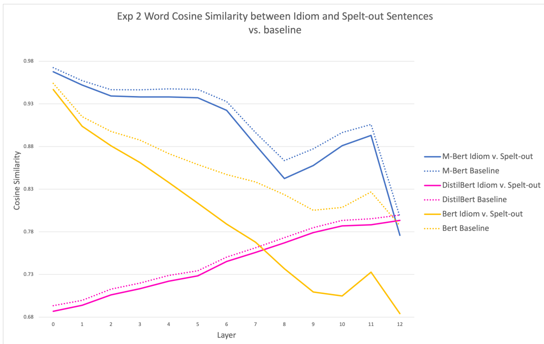

Figure 3: Experiment 2 - Cosine similarities of word embeddings between idiomatic and literal use of the word

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 for the idiom sentence and the two unrelated literal sentences. We calculate the cosine similarity for the word embedding (1) between idiom and literal context 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

[2]We concatenated the activations of all sentence tokens into a single flattened vector[3]. 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.

(dotted line) is higher than between idiom and literal context (solid line). What is surprising is the difference among the 3 LMs. Just like in experiment 1, DistilBERT shows less variations across layers. For BERT, the similarity of word embedding between literal and idiom context drop significant more than between two literal contexts. This confirms our hypothesis that the semantic meaning of idioms are 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. On the other hand, DistilBERT behaves in the opposite way - word embedding actually increases across layers (though overall word embeddings are less similar than BERT and M-BERT). This leads to the question whether the internal structure of DistilBERT - due to its distillation training - is different to LMs trained from language directly.

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 expression functions as a single unit (Sinclair and Sinclair, 1991).

2.4.1 Methods and Results

For each idiom sentence, we select a word inside the idiom and the indices of the target word (e.g. "bush") 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 attention in each layer of LMs from all other sentence tokens to the target word. For all three language models, sentence tokens pays *less* attention to a word inside an idiom (solid lines) than it does to the same word in a literal context (dotted lines), supporting the idea that LMs see idioms as more idiosyncratic units. Once again we observe differences among the three LMs, where DistilBERT shows less variation internally compared to Bert and Multilingual BERT.

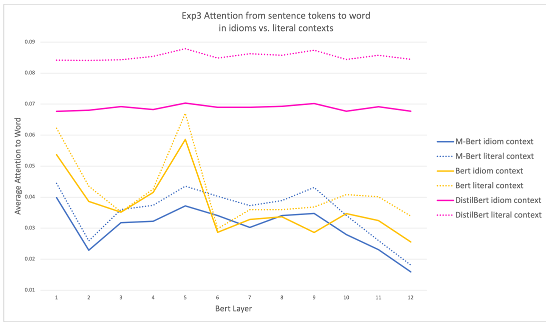

Figure 4: Experiment 3 - Attention from other sentence tokens to word inside an idiom sentence versus a literal sentence

3 Discussion

We investigated how Transformer LMs process idioms across its layers on a sentence level and word level. Experiment 1 shows that on a sentence level, LMs represents 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. The results shed light on the inner workings of LMs on idiom processing. Interestingly, DistilBERT demonstrates less variations across layers compared to Bert and Multilingual BERT, opening the question whether the distillation training of DistilBERT, as opposed to learning from language directly for BERT, reduces internal nuances across layers. We intend to investigate this question in future studies.

4 Conclusion

Idiomatic expressions are part and parcel of everyday language use. This study investigates the inner workings of idiom processing in 3 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. We discovered differences among different LMs, especially between BERT and DistilBERT, raising future questions of the differences in internal structures in different language models.

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)*.

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.

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*.

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.

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. In *International Conference on Text, Speech, and Dialogue*, pages 87–94. Springer.

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

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.