Does BERT Understand Idioms? A Probing-Based Empirical Study of BERT Encodings of Idioms

Minghuan Tan and **Jing Jiang**
School of Computing and Information Systems
Singapore Management University
mhtan.2017@phdcs.smu.edu.sg jingjiang@smu.edu.sg

Abstract

Understanding idioms is important in NLP. In this paper, we study to what extent a pre-trained BERT model is able to encode the meaning of a potentially idiomatic expression (PIE) in a certain context. We make use of a few existing datasets and perform two probing tasks: PIE usage classification and idiom paraphrase identification. Our experiment results suggest that BERT indeed is able to separate the literal and idiomatic usages of a PIE with high accuracy. It is also able to encode the idiomatic meaning of a PIE to some extent.

1 Introduction

Understanding idiomatic expressions is important for NLP tasks such as sentiment analysis (Balahur et al., 2010; Williams et al., 2015) and machine translation (Isabelle et al., 2017; Shao et al., 2018). However, due to the non-compositionality of idioms, it remains a challenge to model the semantic meanings of idioms effectively (Sag et al., 2002; Shwartz and Dagan, 2019).

BERT is a contextualized pre-trained language model that has been widely used and proven to be highly effective for many NLP tasks (Devlin et al., 2019). To better understand how BERT works, recently the community has adopted the approach of *probing*, where a *probing task* is designed to test whether BERT encodings contain sufficient information to perform the task well. Examples of probing tasks include POS tagging and parsing (Hewitt and Liang, 2019; Wu et al., 2020) as well as semantic reasoning tasks such as understanding numbers (Wallace et al., 2019).

It is therefore also natural to ask whether BERT encodes any knowledge about the usage and meanings of idioms, given that BERT was trained on huge corpora, which must contain many idiomatic expressions. However, this problem has not been well explored. To the best of our knowledge, the closest existing work is by Shwartz and Dagan (2019), who studied whether pre-trained (static and contextualized) word embeddings can detect meaning shift and implicit information of phrases, with the help of several probing tasks. However, we believe there is a need for further exploration. We note that Shwartz and Dagan (2019) did not specifically focus on idioms; only one of the six probing tasks was directly related to idioms, and only idiomatic noun compounds were studied. Since English idioms have different syntactic structures, it would be useful to experiment with a higher coverage of different types of idioms.

In this paper, we focus on probing BERT to understand whether BERT embeddings can encode the meanings of a diverse range of different types of idioms. We propose two probing tasks to test whether BERT understands idioms. First, given a context containing a potentially idiomatic expression (PIE), the task is to decide whether the meaning of the PIE is literal or idiomatic, based on the BERT-encoded contextualized embedding of the PIE. We hypothesize that if pre-trained BERT could perform the task well, it would indicate that BERT knows the difference between literal and idiomatic usages of the same expression based on its context. For this task, we use a large dataset recently released by Haagsma et al. (2020), which covers 1756 unique idioms and 50K contextual sentences, much larger and more diverse than the idiomatic noun compounds dataset used by Shwartz and Dagan (2019). However, this task is not sufficient to show whether BERT truly understands the *idiomatic meaning* of a PIE. In order to test this, we design a second probing task based on existing idiom paraphrase datasets. The task is to select the correct paraphrase of an idiom among a set of candidate phrases based on the cosine similarity between the idiom's BERT embedding and these

Proceedings of Recent Advances in Natural Language Processing, pages 1397–1407
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_156

candidate phrases' BERT embeddings. We hypothesize that if the correct paraphrase could be ranked higher than other irrelevant phrases, it would indicate that BERT indeed understands the idiomatic meaning of the idiom.

It is important to note that our objective is not to improve the performance of the two tasks by designing effective learning methods; rather, the objective is to use these two tasks to probe pretrained BERT in order to understand how much BERT encodes the meanings of idioms. Therefore, the models for the two probing tasks are simple models without many parameters to be learned.

Through our empirical study using both the original BERT and ERNIE2 (Sun et al., 2020) (an improved version of BERT), we find that compared with non-contextualized embedding representations of PIEs, contextualized BERT and ERNIE2 embeddings of PIEs can clearly achieve higher accuracy for PIE usage classification, with an accuracy level around 90%, suggesting that BERT can use the context to accurately guess whether an expression is used literally or idiomatically. For paraphrase identification, we find that BERT and ERNIE2 perform significantly better than a random baseline, although the absolute performance is still considered low. Since paraphrase identification is itself challenging, to put things in perspective, we also compare with paraphrase identification for general multi-word expressions (MWEs). Contrary to our expectation, we find that identifying paraphrases for general MWEs does not necessarily fare better than for idioms. Further analysis reveals that this is because BERT contextualization actually hurts paraphrase identification for general MWEs but not so for idioms.

2 Related Work

2.1 Probing Tasks

The notion of *probing* (Ettinger et al., 2016) or a *probing task* (Conneau et al., 2018) refers to the use of a classification problem to reveal whether certain linguistic properties of sentences are captured in the input embedding representations of the sentences fed into the classification model. There have been studies investigating what properties of a sentence its embedding might have contained (Ettinger et al., 2016; Shi et al., 2016; Adi et al., 2017). The properties being probed include semantic roles (Ettinger et al., 2016), negation scopes (Ettinger et al., 2016), constituents (Shi et al., 2016), part-of-speech tags (Shi et al., 2016), sentence lengths (Adi et al., 2017), word orders (Adi et al., 2017), agreement information (Giulianelli et al., 2018) and tense of the main clause (Bacon and Regier, 2018). With the emergence of contextualized embeddings such as BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018a), researchers have also applied probing tasks to word-level contextual representations (Tenney et al., 2019), attention mechanisms (Clark et al., 2019) and syntactic knowledge (Peters et al., 2018b; Hewitt and Manning, 2019). Probing phrasal representations to study lexical composition has also attracted attention. Jawahar et al. (2019) found that the compositional scheme underlying BERT mimics classical, tree-like structures. Shwartz and Dagan (2019) conducted a series of experiments and concluded that lexical composition can shift the meanings of the constituent words and introduce implicit information. Yu and Ettinger (2020) reminded us that phrase representation in transformer models still relies heavily on word content, with little evidence of sophisticated composition of phrase meaning like that done by humans. Our work differs from these existing studies in that we focus on idiomatic expressions rather than any phrases, and we use a recently released large-scale idiom dataset to facilitate our study.

2.2 Potentially Idiomatic Expressions

Potentially Idiomatic Expressions (PIEs) originate from multiword expressions (MWEs) which have both an idiomatic interpretation and a literal interpretation, for example, *spill the beans*. Identifying the correct meaning of a PIE in a certain context is crucial for many downstream tasks including sentiment analysis (Williams et al., 2015), automatic spelling correction (Horbach et al., 2016) and machine translation (Isabelle et al., 2017). There has been both supervised (Sporleder and Li, 2009) and unsupervised (Haagsma et al., 2018; Kurfalı and Östling, 2020) approaches to solve this problem. For example, Feldman and Peng (2013) treated idiom recognition as outlier detection, which does not rely on costly annotated training data. Peng et al. (2014) incorporated the affective hypothesis of idioms to facilitate the identification of idiomatic operations. Different from these studies, our object is not to improve the performance of idiom recognition but rather to use the task as a probing task to understand the capabilities of BERT to encode idioms. With newly created large scale dataset *MAGPIE* (Haagsma et al., 2020), we can further investigate how contextualized word representations works for idiomatic expressions and literal ones.

2.3 Paraphrase Identification

Paraphrase identification aims to determine whether a pair of language units such as sentences have the same meaning (Kauchak and Barzilay, 2006) or whether a given paraphrase candidate can replace a given language unit in its context without changing overall semantic meaning of the text (Yimam et al., 2016). Idiom paraphrasing is a challenging task that has been attracting continuous attention from the community. For example, Liu and Hwa (2016) investigated the effectiveness of a phrasal substitution method to replace idioms with literal expressions, indicating that high quality paraphrasing of idiomatic expressions can be achieved. Yimam et al. (2016) researched a paraphrase-scoring annotation task and showed that the contexts have an impact on the ranking of paraphrases. Haagsma et al. (2018) looks at the literal representation of the PIE's figurative sense (similar to dictionary definitions of an idiom's meaning, which can also be treated as paraphrase) to facilitate potentially idiomatic expression classification. Different from the studies above, in this paper, we focus on understanding whether pre-trained BERT models encode the semantic meanings of idioms, using idiom paraphrase identification as the probing task. We adopt three phrase-level paraphrase datasets for our probing task. Using this task, we probe how contextualization in transformers may affect the semantic relatedness of phrases.

3 Probing Tasks

We design two probing tasks to answer two research questions: (1) Can BERT distinguish the idiomatic usage of a PIE from its literal usage? (2) Can BERT understand the idiomatic meaning of an idiom? Both questions are related to the capabilities of BERT to understand idioms, but the second task is more demanding than the first. The two tasks also share similar objectives as the probing tasks designed by Shwartz and Dagan (2019), which aimed to test whether pre-trained word embeddings can detect the shift of meaning of a phrase from its component words, and whether pre-trained word embeddings understand the implicit meaning of a phrase. However, they are conducting probing at word level, which focuses on whether the meaning of a word in a noun compound (NC) is literal. The dataset only has 90 noun compounds (Reddy et al., 2011). Although they try to augment the dataset using Tratz (2011), the dataset is still limited to 3K. The paraphrase identification task used by them also uses compounds and addresses whether the paraphrase describes the semantic relation between two words of a noun compound (Hendrickx et al., 2013).

In this paper, we use a much larger dataset called MAGPIE (Haagsma et al., 2020) that covers much more potentially idiomatic expressions for phrase-level literal-idiomatic classification. To make the task more challenging, we choose to split the data such that the idiomatic expressions in the training, development, and test sets do not overlap. We further adapt several paraphrase datasets (Liu and Hwa, 2016; Yimam et al., 2016; Pershina et al., 2015) to compare phrasal semantic relatedness for idioms. We compare the effect of BERT encodings at different layers for the two probing tasks to better understand the effect of contextualization.

3.1 PIE Usage Classification

Many MWEs can be interpreted either literally or idiomatically. In some literature, these expressions are defined as potentially idiomatic expressions (PIEs) (Sporleder and Li, 2009; Haagsma et al., 2018, 2020). For example, "spill the beans" can either be used literally to refer to the action of spilling beans or in its idiomatic sense to refer to disclosing some secrete. However, current approaches are investigating this problem with the limitation to one or more syntactic patterns. In this paper, we propose to use the latest large scale dataset MAGPIE to probe how BERT is capturing the difference of literal and non-literal usage of a PIE.

Task Definition. Given a piece of context denoted as $(w_1, w_2, \ldots, w_n)$ containing a PIE with m words, $w_i, \ldots, w_{i+m-1}$, the task is to decide whether the PIE is used with its *literal* meaning or its *idiomatic* meaning. Performance is measured by accuracy. It is important to note that since our goal is to test whether pre-trained BERT can already encode such knowledge, we do not train a classifier *per idiom*. Instead, we train a single binary classifier using a set of training PIEs and their labeled contexts, and test the classifier on a separate set of different test PIEs and their contexts.

Data. We use the *MAGPIE* dataset (Haagsma et al., 2020), which is the largest-to-date corpus of English PIEs and labeled instances of both their literal and idiomatic usages in different contexts. The corpus comprises 1756 unique PIEs and more than 50K contexts, an order of a magnitude larger than previous similar resources. Annotations of *MAGPIE* included various aspects: annotation (dis)agreement, distribution of idiom types, sense distributions across types, composition of the 'other'-category, and influence of genre. An example of *MAGPIE* is given in Table 1. In this paper, we further analyse what might be the reason of BERT's advantage in connection with annotation agreement.

Context: Think of a sunflower turning its flower head towards a source of light — and therefore of energy . The sunflower does not learn by experience to **turn its head** more effectively as it matures , or not to turn at all if it is repeatedly electrically shocked every time it does so .
Annotation: Label: literal PIE: turn head Confidence: 0.75 Genre: W nonAc: nat science Judgment Count: 4 Variant Type: combined-inflection Label Distribution: {'idiomatic': 0.25, 'literal': 0.75}[1]

Table 1: An example from MAGPIE dataset with details of annotations.

3.2 Idiom Paraphrase Identification

In this paper, to further understand whether BERT has learned the idiomatic meaning of phrases, we propose the Idiom Paraphrase Identification probing task to check whether contextualized representations of PIEs encoded by BERT have shifted meanings that are closer to their paraphrases.

Task Definition. Given a piece of context denoted as $(w_1, w_2, \ldots, w_n)$ containing a PIE $w_i, \ldots, w_{i+m-1}$ where the PIE is known to be used idiomatically, and given a set of candidate phrases $\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$, where each $p_l \in \mathcal{P}$ is a MWE and one of them is a paraphrase of the given idiom, the task is to identify the correct paraphrase from $\mathcal{P}$. We cast this task as a ranking problem and use Mean Reciprocal Rank (MRR) to measure the performance.

Data. We combine different resources described below to create the data needed to perform this paraphrase identification task. Specifically, we create three datasets: (1) **Idioms-MWEs**, (2) **MWEs-MWEs**, and (3) **Idioms-Idioms**, Details of the collection of these three datasets are listed below:

- **Idioms-MWEs:** We use the idiom paraphrase dataset created by Liu and Hwa (2016). Each instance in this dataset is a context sentence containing an idiom together with a phrase that can substitute the idiom in the context. The dataset was created by shortening the definitions of these idioms from a dictionary and performing appropriate grammatical and referential transformations to ensure that the idiom substitution fits seamlessly into the original context. The paraphrases have also been verified and refined by human annotators. This gives us a dataset with high quality paraphrases of idiomatic expressions. The dataset contains 171 unique idioms, each with a single context sentence and a paraphrase.

- **MWEs-MWEs:** Since paraphrase identification itself is likely a challenging task even for non-idiomatic MWEs, in order to put things in perspective, we also make use of another paraphrase dataset that contains pairs of MWEs that are paraphrases. Yimam et al. (2016) investigated the impact of context for the paraphrase ranking task using both multi-word expressions and single words. The dataset covers 17k data points (2k MWEs and 15k single word) annotated through crowd-sourcing. The 2k MWEs are of particular interest to us in this probing task. We processed the original dataset by retaining only those paraphrase pairs with a human agreement score of 4, which gives us a final set of 176 entries of a MWE in a context as well as their paraphrases. We find that these 176 entries do not overlap with the PIEs in the MAGPIE dataset, suggesting that these MWEs are likely all non-idiomatic expressions. By performing paraphrase identification on this dataset, we can get a sense of the expected performance for paraphrase identification on phrases that are not idiomatic.

- **Idioms-Idioms:** Pershina et al. (2015) presented idiomatic expressions as a new domain for short-text paraphrase identification and

[1] For the other labels that are not used in this paper, we refer the reader to the original paper for details.

	Size	Example	
		Sentence	Paraphrase
Idioms-Idioms	158	This Cuban Black Bean recipe is pretty much as **easy as beans** get and they are SO delicious.	piece of cake
Idioms-MWEs	171	If only I could **soup up** this computer to run just a little faster.	increase the power of
MWEs-MWEs	176	She constantly complains of boredom as her presence **at home** is merely decorative , while her husband is heavily involved in his scholarly interests .	in her house

Table 2: Paraphrase evaluation datasets. We select one example from each dataset.

released a dataset of 1.4K annotated idiom paraphrase pairs and 2.4K idioms with definitions. However, no context is provided for each idiom. We use this dataset jointly with MAGPIE to construct an evaluation dataset where each entry has an idiom usage label and a definition of the PIE if it is used idiomatically. We use the 91 Idiom-Idiom paraphrase pairs to construct a more challenging split to check if BERT can perceive these paraphrases. By switching the order of each idiom pair, we obtain 192 candidate entries. We retrieve contexts with *idiomatic* label for each idiom pair from MAGPIE to construct the evaluation dataset. For those entries that do not exist in MAGPIE, we retrieve online examples like Wiktionary manually. We filter out some of the entries which share duplicate contexts or have the source idiom being only a naive variation of the target. At the end of the process, we get 158 entries.

For each dataset, we list its size and one example in Table 2.

To create the set of candidate paraphrases, we simply pool the paraphrases of all the entries of the three datasets together as the set of candidate paraphrases for all instances.

4 Experiments

For each of the two probing tasks above, we use pretrained BERT[2] and ERNIE2[3] to process each context $(w_1, w_2, \dots, w_n)$. Following standard practice, we prepend the `[CLS]` token to the beginning of the sequence and append the `[SEP]` token to the end. The sequence is then fed into an L-layer BERT. Let $\mathbf{h}_i^k \in \mathbb{R}^d$ denote the hidden vector produced by the kth layer of BERT representing w_i. When $k = 0$, $\mathbf{h}_i^0$ denotes the combined representation of the word embedding, the position embedding and the token type embedding before it is fed into the transformer-based encoder.

For each PIE, we get a sequence of hidden vectors at the kth layer for the m tokens inside this PIE as follows: $\mathbf{p}^k = (\mathbf{h}_i^k, \mathbf{h}_{i+1}^k, \dots, \mathbf{h}_{i+m-1}^k)$. We will use these contextualized BERT embeddings of the PIE as input to the model for the probing tasks. Note that when training the model for a probing task, BERT is not fine-tuned.

For both probing tasks, we experiment with both the original BERT (Devlin et al., 2019) and ERNIE2 (Sun et al., 2020), which supports phrase masking by using lexical analysis and chunking tools to get the boundary of phrases in the sentences. Our code and data are released on github [4].

4.1 PIE Classification

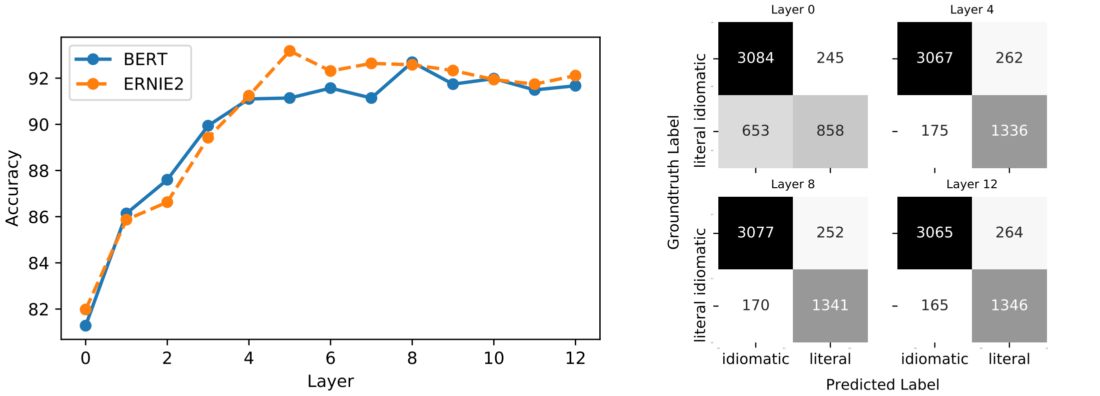
After we get the hidden representation $\mathbf{p}^k = (\mathbf{h}_i^k, \mathbf{h}_{i+1}^k, \dots, \mathbf{h}_{i+m-1}^k)$ of the PIE, we further encode the sequence into a single vector using a bidirectional LSTM encoder. We then treat this vector as input to train the binary PIE usage classifier using a linear classifier.

We show the accuracy of the trained PIE usage classifier on both the development set and the test set in Table 3. We include a baseline BL-majority that always predicts the usage to be idiomatic. This is because we observe that there are more instances in this dataset labeled as idiomatic than literal. We also include another baseline BL-GloVe, which uses the static GloVe word embeddings (Pennington et al., 2014) to replace the BERT encoded representations. For BERT embeddings, we include

[2] huggingface.co/bert-base-uncased

[3] huggingface.co/nghuyong/ernie-2.0-en

[4] https://github.com/VisualJoyce/CiYi



(a) PIE usage classification accuracy on test data with different Transformer layers.

(b) Confusion matrix for Layer-0, Layer-4, Layer-8 and Layer-12.

Figure 1: PIE usage classification.

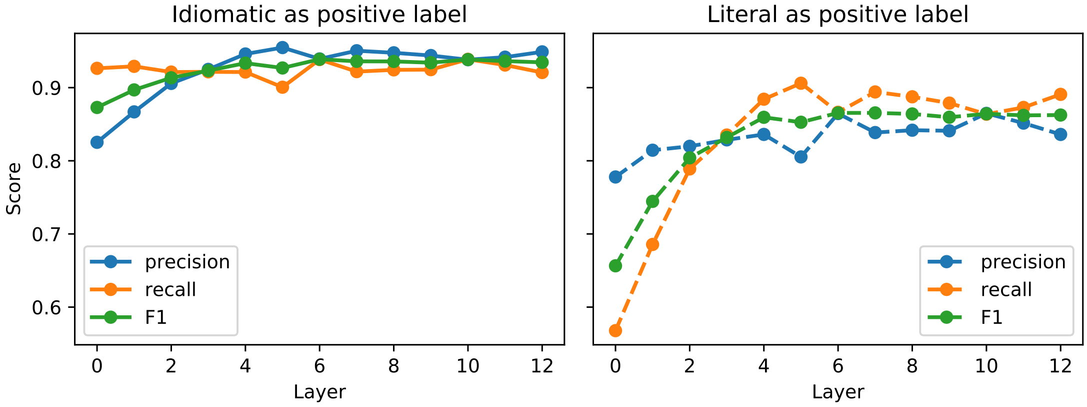


Figure 2: F1 score, precision and recall curve for different layers in BERT. We list both cases that either choosing *idiomatic* or *literal* as the positive label.

the results using the bottom layer (Layer-0) and the results using the final layer (Layer-12). Including Layer-0 is for us to observe how the static embeddings of BERT have performed.

From the table, we can draw the following conclusions: (1) The baseline method **BL-majority** achieves an accuracy above 50%. This shows that the dataset is not balanced, with more instances of idiomatic usage. (2) Using Layer-0 of BERT and ERNIE2, i.e., using only static word embeddings, we can see that the performance is always above 80% and is very close to **BL-GloVe**. This suggests that even the static word embeddings contain some prior knowledge about whether the expression is literal or idiomatic. (3) Using Layer-12 of BERT and ERNIE2, we can see that the accuracy of PIE usage classification significantly increased compared with using Layer-0. In fact the absolute accuracy level is quite high, reaching 90%. This confirms that with BERT contextualization, the embeddings of the PIE better reflect the usage of the PIE, allowing the classifier to easily predict whether the PIE is used literally or idiomatically. This shows that BERT can indeed encode the knowledge about the usage of a PIE.

		Dev	Test
BL-majority		71.76	68.78
BL-GloVe		80.52	82.05
BERT	Layer-0	83.90	81.28
BERT	Layer-12	**90.33**	91.67
ERNIE2	Layer-0	84.65	81.98
ERNIE2	Layer-12	89.03	**92.11**

Table 3: PIE classification accuracy.

Given the large gap between the classification accuracy using Layer-0 and Layer-12, next we experiment with other intermediate layers of the Transformer architecture for BERT and ERNIE2. The results are shown in Figure 1a. From the figure we find that starting from around Layer-4 the performance stabilizes and the last layer is not necessarily the one with the best performance. This shows that BERT requires just a few rounds of contextualization to encode the idiom usage information.

To better understand how BERT contextualization improves PIE usage classification, we further zoom into the two different types of errors: (1) literal usage mistakenly classified as idiomatic usage, and (2) idiomatic usage mistakenly classified as literal usage. We show the numbers of these error cases in four confusion matrices in Figure 1b (one confusion matrix for one of Layer-0, Layer-4, Layer-8 and Layer-12), where the lower-left corner shows the first type of errors and the upper-right corner shows the second type of errors. In Figure 2, we further show the precision, recall and F1 scores across all the layers by either choosing *idiomatic* or *literal* as the positive label. We observe that interestingly the error reductions from Layer-0 to Layer-12 comes mostly from the group *literal-idiomatic* where literal expressions are wrongly predicted to be idiomatic. We hypothesize that this is because without contextualization, some of the words in these PIEs tend to indicate that the PIEs are used idiomatically, probably because these words have appeared often in other idiomatic expressions in the training data; but after considering the specific contexts these PIEs are placed in, i.e., with BERT contextualization, the model recognizes that these contexts are semantically similar to the literal meanings of the tokens inside these PIEs, and therefore predict the usage as being literal. This shows that with more contextualization, BERT embeddings help the most in recognizing literal usages of PIEs.

We further ask the question whether those instances where BERT embeddings did not do well for the PIE usage classification task are those instances where human annotators' agreement is also low. To answer this question, we show the average annotation agreement scores on the test set for correctly predicted instances and incorrectly predicted instances. The statistics are shown in Figure 3. The red line shows the average agreement score over *all* test instances, the green line shows the average agreement score over those instances whose ground truth labels are "idiomatic", and the blue line shows the average agreement score over those instances with the ground truth label "literal". We can see that human annotations have a clearly higher degree of agreement on those idiomatic usages of PIEs, but a lower agreement when a PIE is likely used literally. The four bars in Figure 3 shows the average agreement scores of correctly and incorrectly predicted instances, grouped by the ground truth labels. We can see that clearly those incorrectly predicted instances (shown in light gray bars) have clearly lower human agreement scores compared with the correctly predicted ones. This verifies our hypothesis that the model tends to make mistakes on those instances which humans also find hard.

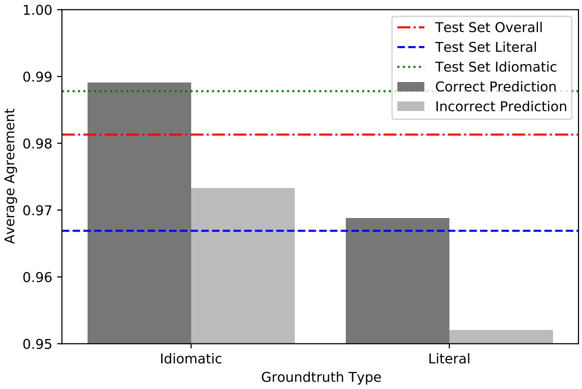


Figure 3: Average agreement score for predictions in Layer-12. Horizontal lines are average annotation agreement scores over test set: (1) Idiomatic cases, (2) Literal cases, (3) Overall.

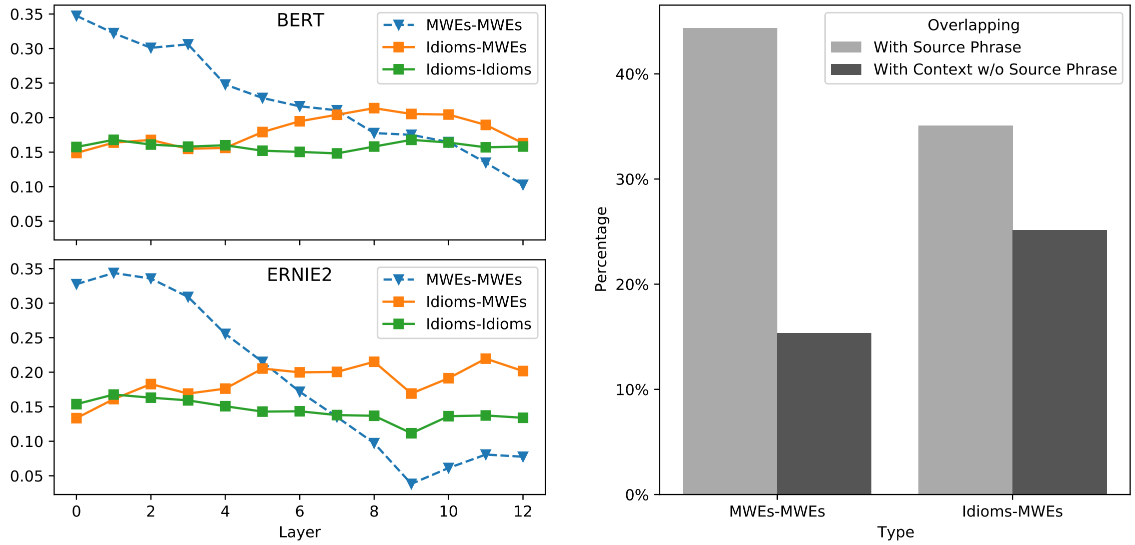
4.2 Paraphrase Identification

	Idioms │ MWEs	MWEs │ MWEs	Idioms │ Idioms
BL-random	0.013	0.013	0.013
BERT	0.163	0.104	0.154
ERNIE2	0.202	0.078	0.136

Table 4: MRR scores for paraphrase ranking.

For the paraphrase identification task, after we get the hidden representation $\mathbf{p}^k$ of the PIE in its context, we take the average of these vectors to obtain a single vector. For each candidate paraphrase, we perform the same encoding, without any context, and then take the average of the produced hidden vectors. Finally, we rank the candidates based on the cosine similarity between the PIE's embedding and the candidate's embedding.

The Mean Reciprocal Rank (MRR) scores are



(a) Paraphrase ranking MRR across layers for different splits.

(b) Percentage of pairs with word overlap.

Figure 4: Paraphrase identification.

listed in Table 4. For comparison, we consider a baseline that randomly ranks the candidates. We can observe the following from the table: (1) BERT and ERNIE2 can perform better than the random baseline on Idioms-MWEs, although the absolute values of MRR are low. This shows that BERT contextualized embeddings can still encode the idiomatic meanings of idioms to some extent. (2) We also observe that identifying paraphrases for general multi-word expressions (MWEs-MWEs), which are likely not idiomatic, is not easier than for idioms. This is counter-intuitive and we will show further investigation below. (3) Identifying paraphrase idioms of idioms (Idioms-Idioms) is a bit harder than identifying general multi-word-expression-based paraphrases. This maybe because the candidate idioms are not contextualized, and therefore their embeddings do not reflect their idiomatic meanings.

To better understand why paraphrase identification for general MWEs has even lower performance than for idioms, we again test the performance using different layers of BERT/ERNIE2 embeddings. The results are shown in Figure 4a. Now it is clear that with non-contextualized embeddings (i.e., Layer-0), paraphrase identification for general MWEs is actually much easier than for idioms. This is intuitive because the meaning of non-idiomatic MWEs can be derived from their component words and therefore contextualization is not needed. The figure also shows that with more contextualization, performance of paraphrase identification for general MWEs is largely hurt, but this is not the case for idioms. It's also interesting that, for Idioms-Idioms, the MRR scores do not change much with layers. We think this may due to both an idiom and its idiomatic paraphrase share less overlap with the context.

Noticing that the performance of paraphrase identification for **Idioms-MWEs** surpasses **MWEs-MWEs** at Layer-8, i.e., when there is some degree of contextualization, we conduct some further analysis to understand why. Specifically, given a query idiom (or query MWE) q, its context c, and its ground truth paraphrase MWE p, we would like to check if p tends to have common words with q and c, respectively. Our hypothesis is that if p shares common words with c, then contextualized word embeddings are helpful because they encode the context c. We show our analysis in Figure 4b. In the left hand side of the figure, the light gray bar shows the percentage of test instances in the MWEs-MWEs dataset where the query MWE q shares at lease one common word with the ground truth paraphrase p, and the dark gray bar shows the percentage of test instances in MWEs-MWEs where the context c shares at least one common word with the ground truth paraphrase p. The right hand side of the figure shows the same percentages for the Idioms-MWEs dataset. We can see that for MWE-MWE paraphrase pairs, it is less common for the ground truth paraphrase to share a common word with the context of the query phrase, compared with Idiom-MWE paraphrase pairs. This is reasonable because for an idiom, its idiomatic meaning is often not directly linked to the semantic meanings of their component words, and therefore words in the idiom itself may not overlap with words in its paraphrase; on the other hand, the context where an idiom appears may imply the idiom's idiomatic meaning, and therefore may have word overlap with the paraphrase. The statistics shown in Figure 4b shows that because for MWEs, their paraphrases are less likely to share common words with the contexts where the MWEs appear, contextualization done by BERT therefore not only is not so useful but also may harm the performance of paraphrase identification.

5 Conclusion

In this paper, we use two probing tasks to study whether BERT understands English idioms. In conclusion, we find that BERT is able to detect idiomatic usages of a PIE with a high accuracy, and with more contextualization as the layer increases, BERT helps the most in recognizing literal usages of PIEs. However, this only proves that BERT is effective in detecting meaning shift for idiomatic expressions. To further probe if the shifted meanings are closer to their paraphrases, we adopt the paraphrase identification task by gathering three different types of paraphrase pairs, MWEs-MWEs, MWEs-Idioms and Idioms-Idioms. Our experiments show that BERT is able to encode the idiomatic meaning to some extent. However, contextualization may have different effects for MWEs and idioms, which still requires further exploration to fully explain.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *international conference on learning representations*.

Geoff Bacon and Terry Regier. 2018. Probing sentence embeddings for structure-dependent tense. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 334–336, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *BlackBoxNLP@ACL*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *Computational Linguistics and Intelligent Text Processing*, pages 435–446, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2018. The other side of the coin: Unsupervised disambiguation of potentially idiomatic expressions by contrasting senses. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 178–184, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. 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*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94, online. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, Thomas R. McCoy, Najoung Kim, Van Benjamin Durme, R. Samuel Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *ICLR*.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. University of Southern California.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375 – 7385.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl, and Chris Biemann. 2016. Learning paraphrasing for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 1–10, Berlin, Germany. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.