Finding BERT's Idiomatic Key

Vasudevan Nedumpozhimana and John D. Kelleher
ADAPT Research Center
Technological University Dublin
Dublin, Ireland
{Vasudevan.Nedumpozhimana,john.d.kelleher}@tudublin.ie

Abstract

Sentence embeddings encode information relating to the usage of idioms in a sentence. This paper reports a set of experiments that combine a probing methodology with input masking to analyse where in a sentence this idiomatic information is taken from, and what form it takes. Our results indicate that BERT's idiomatic key is primarily found within an idiomatic expression, but also draws on information from the surrounding context. Also, BERT can distinguish between the disruption in a sentence caused by words missing and the incongruity caused by idiomatic usage.

1 Introduction

Idioms occur in almost all languages, however the processing of idioms by NLP systems remains extremely challenging (Villavicencio et al., 2005; Sporleder and Li, 2009; Salton et al., 2014). One reason for this is that many expressions can be used both literally or idiomatically. Fazly et al. (2009) distinguish between identifying whether an expression has an idiomatic sense (*idiom type* classification) and identifying whether a particular usage of an expression is idiomatic (*idiom token* classification), and focus their work on analysing the canonical form (lexical and syntactic) of idiomatic expressions. The related work on *idiom token* classification at a sentence level includes (Sporleder and Li, 2009; Li and Sporleder, 2010a,b; Peng and Feldman, 2017; Fazly et al., 2009; Salton et al., 2016, 2017). Of particular relevance is Salton et al. (2016) which demonstrated that it is possible to train a generic (as distinct to expression specific) idiom token classifier using distributed sentence embeddings. Of note here is that Salton et al. (2016) used Skip-Thought vectors rather than the more recent contextual embeddings such as BERT (Devlin et al., 2019) and also that these results indicate that language model based embeddings encode information from a sentence relating to the literal or idiomatic usage of expressions.

Cacciari and Tabossi (1988) proposed one of the most accepted psycholinguistic theories of how humans identify the presence of an idiom. This theory posits that there is a part of every idiomatic expression that must be processed (i.e., accessed from the mental lexicon) before the idiomatic meaning of the expression can be recognised. This special part of an idiomatic expression is known as the *idiomatic key*. The theory leaves open how incongruency between an expression and the context it occurs within might trigger a figurative interpretation.

Given the empirical results of Salton et al. (2016) and the psycholinguistic work of Cacciari and Tabossi (1988) one question that arises is where in a sentence is the *idiomatic key* for models such as BERT: is it predominantly local to the expression or not? Note, that here we are using a broader concept of idiomatic key than that proposed by Cacciari and Tabossi (1988): they limit the idiomatic key to be a part of an expression, whereas we use the concept of idiomatic key to be the part of a sentence that provides BERT with a signal that an expression is being using idiomatically. Answering the question of where BERT's idiomatic key is can provide insight into how BERT, and similar systems work, and also into human language processing. In this paper we address this question by using a probing style experiment (Conneau et al., 2018) combined with various input masking techniques.

Section 2 describes the dataset, embeddings, and model types that we use. Section 3 reports baseline experiments that examine the strength of the idiomatic usage signal encoded in BERT embeddings, and Section 4 reports a second set of experiments where various masking techniques are used to analyse where in a sentence BERT's *idiomatic key* is located. Section 5 sets out our conclusions.

Proceedings of the 17th Workshop on Multiword Expressions, pages 57–62
Bangkok, Thailand (online), August 6, 2021. ©2021 Association for Computational Linguistics

2 Data, Embeddings, and Models

A probing experiment tests for the presence of information relating to a linguistic phenomenon within an embedding. The methodology involves using the embedding as the input to a model that is trained to predict whether the linguistic phenomenon is present in the original linguistic input or not. If the model can achieve a high-accuracy on the task this is taken as evidence that the embedding encodes information on the linguistic phenomenon. Indeed, the work by Salton et al. (2016) is an early example of probing, in that instance probing Skip-Thought vectors and idiom token classification.

For our experiments we used the VNIC data set (Cook et al., 2008). The VNIC dataset contains 2984 sentences across 56 idiomatic expressions. Each sentence contains one of the target expressions and is labelled as: idiomatic, literal or unknown usage. Of these 2984 sentences 2016 sentences are used in idiomatic sense, 550 sentences are used in literal sense, and remaining sentences are labelled as unknown. A model trained on such an imbalanced dataset will likely be biased towards the majority class label (in this case the idiomatic label) and such a bias would be a confounding factor in our masking experiments. In our experimental setup the signal we use to identify BERT's idiomatic key is how the ablation of different types of information (via various forms of masking) affects the likelihood BERT returns for idiomatic usage within a sentence. If BERT is biased towards idiomatic usage based on class distribution untangling the effects of this bias from the effects of information ablation would make our analysis much more complex. To control for this bias we downsampled the dataset to make sure that the dataset has a balanced label distribution. We selected all 550 sentences with literal usage and 550 sentences with idiomatic usage by randomly down sampling 2016 idiomatic sentences for our probing experiment. We repeated the down sampling of idiomatic sentences 20 times to prepare 20 different versions of the dataset, and for each version of the down sampled dataset we then split the 1100 sentences into a training set with 80% of samples and a testing set with the remaining 20% of samples with stratified label distribution. Consequently, downsampling not only enables us to balance the class labels but also gives this opportunity to repeat experiment with many versions of dataset and this provides the benefit of cross validation. For each experiment we have run the experiment independently on each of the 20 down sampled versions of the dataset, and then calculated the macro average score across these 20 independent runs.

For each down sampled version of the dataset we used a bert-base-uncased pretrained BERT model[1] to generate sentence embeddings (Devlin et al., 2019). We use this version of BERT as a representative of BERT based (transformer based) language model family. In this experiment our focus is to analyse the pretained BERT model, and the information signals it uses for the task of idiom token identification, rather than to extend the current state of the art performance on this task and therefore we didn't fine tune the BERT model. This BERT model gives 12 layers of 768 dimensional embeddings for each word in a sentence. We used the average of the final layer of word embeddings as the sentence embedding.

For our probing experiments we trained a multilayer perceptron (MLP) on the training split of each dataset to predict a high probability for embeddings of idiomatic usage sentences and low probability for embeddings of literal usage sentences. The MLP with 768 inputs, one hidden layer of 100 ReLUs, and a logistic unit output layer was implemented using Scikit-learn library (Pedregosa et al., 2011). The MLP was trained using an Adam solver (Kingma and Ba, 2014) using the Scikit-learn default hyper parameters and a convergence criterion of 200 epochs. We define the probability score of a sentence predicted by the trained MLP model as the score of idiomaticity of that sentence.

3 Baseline Results

To evaluate the MLP we use the mean idiomaticity scores on the idiomatic and literal segments of the test sets, where the ideal score of an idiomatic sentence is 1.00 and a literal sentence is 0.00. Consequently, the closer the average score returned by the model on idiomatic sentences is to 1.00 the stronger the model, and similarly the closer the average score returned by the model on literal sentences is to 0.00 the stronger the model. The *Baseline* scores in Table 1 show the average scores returned by the models on the idiomatic and literal segments of the test sets. The MLPs have good performance on both idiomatic sentences (0.85 against the ideal 1.00) and on literal sentences (0.17 against

[1] 12-layer, 768-hidden, 12-heads, and 110M parameters, trained on lower-cased English text

the ideal 0.00). This strong performance indicates that the MLPs effectively predict the idiomaticity of both idiomatic and literal sentences, and furthermore that BERT sentence embeddings encode information relating to idiomatic usage.

4 Masking Experiments

Our primary objective is to locate where BERT's idiomatic key is located within a sentence, is it concentrated within the expression or not. In order to gather information on this we conducted an experiment to test how the idiomaticity scores returned by the MLP model changed when we masked different parts of the input. The intuition behind our experimental design is that if we mask the components of a sentence that are informative regarding idiomatic usage within the sentence this should result in the MLP model shifting their scores for a sentence towards 0.5 in an amount that is proportional to the informativeness of the masked component, because the model will have less certainty regarding the idiomatic, or literal, usage within the sentence. Note, that the test sets used in these masking experiments are the same 20 test sets that were used in the baseline experiments. Furthermore, the MLP model tested on each test set is the same model trained using the corresponding training split for the baseline experiment (i.e., the training set is not masked). Consequently, the baseline results discussed above are for the same models used in this experiment.

For this experiment a natural part of a sentence to mask is the expression whose idiomatic usage within the sentence is being assessed. However, given that the idiomatic key may be located outside the target expression we also need to select other components of sentences to be masked. There are many ways we could have selected these components. However, all the target expressions in our data contain two words, a verb and a noun, and so for each sentence we randomly selected two other words for masking. This method has the advantages of simplicity and also matching the number of words masked in the sentence when masking an expression or masking outside the expression.

As a measure for the informativeness of a component (target expression or random selection) with respect to idiomatic usage within the sentence we define *differential idiomaticity* as the difference in idiomaticity score returned by the MLP model for the sentence embedding when the component is present in the input and when it is masked. Our models are trained to score idiomatic usage sentences close to 1.00 and so we expect that for idiomatic usage sentences differential idiomaticity will be positive (between 0.00 and 1.00) because masking part of the input will likely shift the model score towards 0 and the difference between the score for the unmasked input and the masked input will then be positive. Conversely, for literal sentences we expect that differential idiomaticity will be a negative (between 0.00 and -1.00). Overall, the informativeness of a component with respect to idiomatic usage in a sentence is captured by the magnitude of its differential idiomaticity.

We followed two strategies for masking information in a sentence: *word masking* and *embedding masking*. In the word masking strategy, we replace the words in a sentence to be masked using the same [MASK] token as that used by Devlin et al. (2019). Our word masking strategy completely blocks the information from masked words. However, the resulting sentence may not be a valid sentence. Consequently, we also tested a second masking strategy that retained the words in the sentence input into BERT but masked the word embeddings prior to calculating the sentence embedding. We generate the sentence embeddings by taking the average of the final layer of BERT embeddings of all words in the sentence. However, when we apply embedding masking we don't include the final layer embeddings of the words to be masked in the calculation of the sentence embedding.

5 Results and Conclusions

Table 1 presents the average idiomaticity and the differential idiomaticity with respect to Baseline along with p-values from the experiment broken down by component being masked (target expression or random words) and the type of sentence (idiomatic or literal usage) by using the trained MLP model. As noted in the preceding section, we consider the absolute value of differential idiomaticity as an indication of idiomatic information in a component.

For idiomatic sentences we observe that using a word masking strategy masking either the target expression or random words outside of the expression resulted in a statistically significant difference in idiomaticity scores compared with the baseline results (the differential idiomaticity of 0.02 for random word masking has a p-value of 0.026 and the differential idiomaticity of 0.06 for masking the

Masking	Idiomatic			Literal		
	Id	DId	p-value	Id	DId	p-value
Baseline	0.85	-	-	0.17	-	-
Target Expn + Word Mask	0.79	0.06	1.12E-05	0.24	-0.08	2.83E-07
Target Expn + Emb Mask	0.83	0.02	1.91E-11	0.19	-0.02	4.07E-16
Rand Word + Word Mask	0.83	0.02	0.026	0.17	0.00	0.854
Rand Word + Emb Mask	0.85	0.00	0.313	0.17	0.00	0.378

Table 1: Mean Idiomaticities (Id) and Mean Differential Idiomaticities (DId) and p-values

target expression has a p-value of $1.12E{-}05$). The fact that masking the words in the target expression has a larger effect on idiomaticity compared with masking random words outside the expression indicates that the idiomatic key is primarily concentrated within the target expressions, which makes intuitive sense. However, the fact that the differential idiomaticity for random word masking is also statistically significant indicates that for BERT the idiomatic key is not restricted to be within the target expression, but may also occur in in the context. Finally, the fact that word masking has a larger impact on idiomaticity compared with embedding masking suggests that the idiomatic key is not equivalent to a disruption of any type in the sentence, we will return to this below.

For literal sentences, masking of target expression resulted in a statistically significant difference in idiomaticity (the mean differential idiomaticity of -0.08 with word masking has a p-value of $2.83E{-}07$ and the mean differential idiomaticity of -0.02 with embedding masking has a p-value of $4.07E{-}16$), but masking of random words outside target expression shows insignificant difference with both word masking and embedding masking approaches (negligibly small mean differential idiomaticity with word and embedding masking having p-values 0.854 and 0.378 respectively). These results generally mirror the results on idiomatic sentences and suggest that the signal BERT uses to distinguish literal from idiomatic usages of an expression is primarily found in the expression itself.

One question that arises is whether these differential idiomaticity scores actually relate to the removal of specific information relating to idiomatic usage from an embedding or just reflect disruption within the sentence. The signal encoded in an embedding for idiomatic usage within a sentence may, in fact, be some form of high-perplexity or incongruity in the sentence, and so it is very difficult to disentangle different forms of disruption within a sentence: how should we disentangle the surprise of an unexpected word from the surprise of a missing word? Indeed, it may be that by introducing some particular form of disruption (via masking) into a BERT sentence embedding we are in fact simulating an idiomatic key.

The differential idiomaticity scores for the embedding masking is a potential source of information relevant to this topic. The fact that the differential idiomaticity scores resulting from embedding masking are smaller than those generated by word masking reflects the fact that the self-attention mechanism within the BERT architecture means that the final layer embedding for a word encodes information from other words in the sentence. Consequently, the final sentence embedding generated under embedding masking indirectly encodes the information from the masked embeddings (because the unmasked embeddings that are included encode information about the words corresponding to the masked embeddings) and as a result the sentence embedding is less disrupted by the masking process. In other words, the missing word effect is not as strong under embedding masking but the word incongruity effects caused by idiomatic usage could still be present. Given this, the weak differential idiomaticity scores generated using embedding masking might indicate that BERT is able to encode word incongruity within a sentence embedding even if the embedding for the word itself is not included in final calculation of the sentence embedding, and consequently the idiom token classifiers are still able to confidently predict idiomatic usage. More generally, it suggest that BERT embeddings distinguish between the disruption caused by missing words and the type of incongruity introduced into a sentence by the idiomatic usage of an expression.

Another factor to consider here is that in our dataset the target expressions are verb noun compounds. Consequently, these expressions are made up of content words that likely contain topical information. Our experiment shows significant differences in idiomaticity on both idiomatic and literal sentences after masking the target expression. The rise in idiomaticity in literal sentences due to target expression masking might be because of the incongruity caused by the absence of content words in the target expression. Similarly the reduction of idiomaticity in idiomatic sentences after the target expression masking might be because of the reduced incongruity within the sentence caused by the absence of an idiomatic target expression. This suggests that the incongruity caused by presence or absence of a target expression, or other content words, which have topical information might be the idiomatic key of BERT and further experiments are needed to investigate this.

In conclusion, our results indicate that BERT's idiomatic key is primarily found within an idiomatic expression itself, but also relies on some information from the surrounding context. Also, BERT can distinguish between the disruption in a sentence caused by words missing and the incongruity introduced by idiomatic usage. Further investigation regarding the idiomatic information in the surrounding context (for example, by masking different categories of words, such as content words, topical key words, or words with different part of speech categorization) is proposed for future research.

Acknowledgments

This work was partly supported by the ADAPT Centre which is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Funds.

References

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27:668–683.

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.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

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.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Comput. Linguist.*, 35(1):61–103.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Linlin Li and Caroline Sporleder. 2010a. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 683–691, Stroudsburg, PA, USA. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2010b. Using Gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, Los Angeles, California. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jing Peng and Anna Feldman. 2017. Automatic idiom recognition with word embeddings. In *Information Management and Big Data - 2nd Annual International Symposium, SIMBig 2015 and 3rd Annual International Symposium, SIMBig 2016, Revised Selected Papers*, Communications in Computer and Information Science, pages 17–29. Springer Verlag.

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.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2017. Idiom type identification with smoothed lexical features and a maximum margin classifier. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 642–651, Varna, Bulgaria. INCOMA Ltd.

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 Association for Computational Linguistics*, EACL '09, pages 754–762, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comput. Speech Lang.*, 19(4):365–377.