Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions*

Reyhaneh Hashempour
University of Essex
rh18456@essex.ac.uk

Aline Villavicencio
The University of Sheffield
a.villavicencio@sheffield.ac.uk

Abstract

The majority of studies on detecting idiomatic expressions have focused on discovering potentially idiomatic expressions overlooking the context. However, many idioms like ***blow the whistle*** could be interpreted idiomatically or literally depending on the context. In this work, we leverage the Idiom Principle (Sinclair et al., 1991) and contextualized word embeddings (CWEs), focusing on Context2Vec (Melamud et al., 2016) and BERT (Devlin et al., 2019) to distinguish between literal and idiomatic senses of such expressions in context. We also experiment with a non-contextualized word embedding baseline, in this case Word2Vec (Mikolov et al., 2013) and compare its performance with that of CWEs. The results show that CWEs outperform the non-CWEs, especially when the Idiom Principle is applied, as it improves the results by 6%. We further show that the Context2Vec model, trained based on Idiom Principle, can place potentially idiomatic expressions into distinct 'sense' (idiomatic/literal) regions of the embedding space, whereas Word2Vec and BERT seem to lack this capacity. The model is also capable of producing suitable substitutes for ambiguous expressions in context which is promising for downstream tasks like text simplification.

1 Introduction

The task of determining whether a sequence of words (a Multiword Expression - MWE) is idiomatic has received lots of attention (Fazly and Stevenson, 2006; Cook et al., 2007). Especially for MWE type idiomaticity identification (Constant et al., 2017), where the goal is to decide if an MWE can be idiomatic regardless of context, high agreement with human judgments has been achieved, for instance, for compound nouns (Reddy et al., 2011; Cordeiro et al., 2016). However, as this task does not take context into account, these techniques have limited success in the case of ambiguous MWEs where the same expression can be literal or idiomatic depending on a particular context. For example, such models would always classify ***hit the road*** as idiomatic (or conversely always as literal) while the expression could be idiomatic in one context and literal in another. As a consequence, for practical NLP tasks, especially Machine Translation and Information Retrieval, token idiomaticity identification is needed, with the classification of a potential idioms as literal (or idiomatic) in context. For example, ***hit the road*** must be translated differently in "The bullets were ***hitting the road*** and I could see them coming towards me a lot faster than I was able to reverse", and "The Ulster Society are about to ***hit the road*** on one of their magical history tours" (Burnard, 2000).

We argue that successful classification of potentially idiomatic expressions as idiomatic/literal is not possible without taking the context into account. Recently introduced Contextualized Word Embeddings (CWEs) are ideal for this task as they can provide different embeddings for each instance of the same word type. CWEs such as Context2Vec (Melamud et al., 2016) and BERT (Devlin et al., 2019) proved successful in the task of Word Sense Disambiguation (WSD) (Huang et al., 2019; Hadiwinoto et al., 2019). We also argue that disambiguation of potentially idiomatic expressions is analogous to WSD in a sense that it also tries to assign the most appropriate sense to an idiom, i.e. literal, or idiomatic depending on its respective context.

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 72–80
Barcelona, Spain (Online), December 12, 2020

Moreover, we hypothesize that in order to fully exploit the capacity of CWEs, an idiom should be treated as a single token both in training and testing. This hypothesis is inspired by the evidence from psycholinguistic studies which support the idea that the idiomatic expressions are stored and retrieved as a whole from memory at the time of use (Siyanova-Chanturia and Martinez, 2014). It is also rooted in the idea that different types of information are captured in vectors depending on the type of input, i.e. word, character, phrase to the model (Schick and Schütze, 2019). Moreover, this method proved successful in other tasks. For instance, Carpuat and Diab (2010) conducted a study for integrating MWEs in Statistical Machine Translation (SMT) and improved the BLEU score by treating MWEs as single tokens in training and testing.

Contribution: We show that CWEs can be utilized directly to detect idiomaticity in potentially idiomatic expressions due to their nature of providing distinct vectors for the same expression depending on its context. We also apply the Idiom Principle (Sinclair et al., 1991) when training the models which improves the results as expected and supports our hypothesis that an MWE should be treated as a single token both in training and testing the models. We further show that Context2Vec trained based on Idiom Principle is able to provide suitable replacement for MWEs in both the idiomatic and literal senses. To the best of our knowledge, this is the first attempt to integrate Idiom Principle into CWEs and directly use them for the task of identification idiomaticity in MWEs at the token level.

2 Related work

Distributional Semantic Models (DSM) are computational models based on the Distributional Hypothesis (Harris, 1954) and the idea that words occurring in similar contexts tend to have a similar meaning. Recently, two flavours of Distributional Models have been introduced and utilized which are known as contextualized and non-contextualized embedding models. The former produces different embeddings for a word depending on the context and the latter offers only one embedding for a word regardless of the context. Researchers have leveraged DSMs along with linguistic knowledge to deal with identifying MWEs at type (Cook et al., 2007; Cordeiro et al., 2016; Nandakumar et al., 2019) and token level (King and Cook, 2018; Rohanian et al., 2020).

For instance, the degree of linguistic fixedness was used as the basis for Fazly and Stevenson (2006) to apply an unsupervised method to distinguish between idiomatic and literal tokens of verb-noun combinations (VNCs). They argue that idiomatic VNCs come in fixed syntactic forms in terms of passivation, determiner, and noun pluralization. They extracted these forms using known idiomatic/literal VNC patterns and among all variations they determined which were the canonical form(s). Then they classified new tokens as idiomatic if they appeared in their canonical forms.

Cook et al. (2007) leveraged the idea of canonical forms and the Distributional Hypothesis and built co-occurrence vectors representing the idiomatic and literal meaning of each expression based on their context and (canonical) forms. The problem with this model is relying solely on the canonical form to label an expression as idiomatic/literal which is not enough as there are many MWEs, e.g. ***kick the bucket*** that can be in their canonical form and yet have a literal meaning depending on the context they appear in. Hence, each MWE should be disambiguated in its own individual context.

Cordeiro et al. (2016) also built their work based on Distributional Hypothesis and the Principle of Compositionality to classify MWEs as idiomatic/literal. Their idiomaticity identification model at the type level works well for MWEs that are either idiomatic or literal but falls short for idiomaticity identification at the token level when the MWE is ambiguous.

Nandakumar et al. (2019) used different types of contextualized and non-contextualized word embeddings from character-level to word-level models to investigate the capability of such models in detecting nuances of non-compositionality in MWEs. When evaluating the models, they considered the MWEs out of their context which is problematic especially in case of utilizing CWEs as the reason behind the success (Peters et al., 2018; Devlin et al., 2019; Akbik et al., 2019) of these models is in their ability to produce context-specific embeddings for each token.

The main drawback of above-mentioned works is that they do not take the context of each individual expression into account when classifying them. However, there have been some attempts to detect idiomaticity in MWEs in context (at token level) using Distributional Models.

Peng et al. (2015) exploited contextual information captured in word embeddings to automatically recognize idiomatic tokens. They calculate the inner product of the embeddings of the context words with the embedding of target expression. They argue that since the literal forms can predict the local context better, their inner product with context words is larger than that of idiomatic ones, hence they tell apart literals from idiomatic forms.

Salton et al. (2016) exploited Skip-Thought Vectors (Kiros et al., 2015) to represent the sentential context of an MWE and used SVM and K-Nearest Neighbours to classify MWEs as idiomatic or literal in their context. They compared their work against a topic model representation that include the full paragraph as the context and showed competitive results.

King and Cook (2018) proposed a model based on distributed representations, non-CWE to classify VNC usages as idiomatic/literal. First, they represented the context as the average embeddings of context words and trained a Support Vector Machines (SVM) classifier on top of that. They further showed that incorporating the information about the expressions canonical forms boosted the performance of their model.

A related task of metaphor token detection has seen successful results with the combination of CWEs and non-CWEs, along with linguistic features (Gao et al., 2018; Mao et al., 2019). For instance, Gao et al. (2018) used Word2Vec and ELMo (Peters et al., 2018) as embeddings, with a bidirectional LSTM to encode sentences, and a feed-forward neural network for classifying them as literal or metaphoric.

Rohanian et al. (2020) presented a neural model and BERT, to classify metaphorical verbs in their sentential context using information from the dependency parse tree and annotations for verbal MWEs. They showed that incorporating the knowledge of MWEs can enhance the performance of a metaphor classification model.

We follow the intuition that CWEs can be directly used for the task of token level identification of idiomaticity in MWEs due to their ability to produce different embeddings for the different tokens of the same MWE. Our work is also inspired by the Idiom Principle which explains how human distinguish idiomatic expressions.

3 Distributional Models and Idiom Principle

In this work, we use Word2Vec as non-CWEs and leverage the Context2Vec and BERT as CWEs in combination with the Idiom Principle to detect idiomaticity in potentially idiomatic expressions. The embedding models and Idiom Principle are briefly described here.

3.1 Word2Vec

For Word2Vec we use CBOW (Mikolov et al., 2013) which represents the context around a target word as a simple average of the embeddings of the context words in a window around it. For example, for the window size of two, two words before and two words after the target word are considered as the context of the target word whose embeddings are averaged to represent context embeddings. To train our Word2Vec model, we use Gensim (Řehůřek and Sojka, 2010) with window size of 5 and 300 dimensions. We ignore all words that occur less than fifteen times in the training corpus. We perform negative sampling and set the number of training epochs to five as in King and Cook (2018).

3.2 Context2Vec

Context2Vec (Melamud et al., 2016) uses a bidirectional LSTM recurrent neural network, where one LSTM is fed with with the sentence words from left to right, and the other from right to left. Then right-to-left and left-to-right context embeddings are concatenated and fed into a multi-layer perceptron to capture dependencies between the two sides of the context. We consider the output of this layer as the embedding of the entire joint sentential context around the target word. This is a better representation of the context compared to that of Word2Vec, as it takes the order of words into account. To train our Context2Vec model, we use the code provided by the authors[1] having the same configuration for the

[1] https://github.com/orenmel/context2vec

hyper-parameters.

3.3 BERT

Contrary to the Context2Vec, BERT (Devlin et al., 2019) does not rely on the merging of two uni-directional recurrent language models, but using the transformer (Vaswani et al., 2017) encoder, it reads the entire sequence of words at once. It also benefits from the next sentence prediction feature which helps capture more contextual information. To train our BERT model, we use BERT-Base keeping the configuration of the hyper-parameters intact.

3.4 Idiom Principle

The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments (Sinclair et al., 1991). In other words, MWEs are treated as single tokens in mental lexicon when stored in or retrieved from memory. One of the highly cited definitions of MWEs is also supports the Idiom Principle; Wray (2002) defines MWEs as a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated:that is, stored, retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

Model	Idiomatic			Literal			Ave.F
	P	R	F	P	R	F	
Original Models							
Word2Vec	0.75	0.80	0.77	0.51	0.51	0.51	0.64
Context2Vec	0.76	0.78	0.77	0.62	0.61	0.61	0.70
BERT	0.80	0.81	0.80	0.60	0.61	0.60	**0.71**
Idiom-Principle-inspired							
Word2Vec	0.70	0.73	0.72	0.56	0.60	0.58	0.65
Context2Vec	0.80	0.82	0.81	0.71	0.72	0.71	**0.76**
BERT	0.77	0.79	0.78	0.66	0.63	0.64	0.71
King and Cook (2018)							
W2V-CF	0.82	0.88	0.83	0.63	0.54	0.56	0.69
W2V+CF	0.83	0.89	0.85	0.76	0.68	0.69	**0.77**

Table 1: Precision (P), recall (R), and F1 score (F), for the idiomatic and literal classes, as well as average F1 score (Ave.F) for the original and the Idiom-Principle-Inspired models. The results of King and Cook (2018) are also reported for comparison.

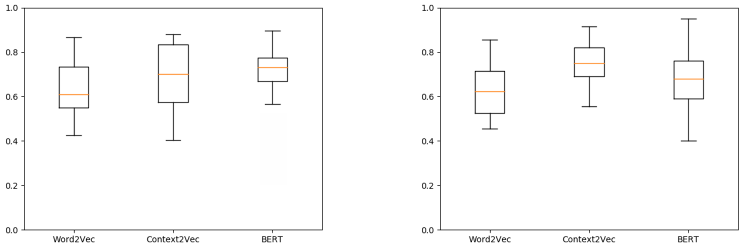


Figure 1

Figure 2

Table 2: Box plot for average F-score for the original (Figure 1) and Idiom-Principle-inspired (Figure 2) models.

4 Experimental Setup

To test the hypothesis, we build 6 different context representations using three embedding models: Word2Vec, Context2Vec and BERT in two different settings: 1- Their original models where each expression is not treated as a single token 2- Our own models, which we call Idiom-Principle-Inspired models, where each expression is treated as a single token.

For the first setting, we use the original pre-trained models, Word2Vec, Context2Vec and BERT-base-uncased. For the second setting, we use BNC corpus (Burnard, 2000) and first lemmatize it using spaCy (Honnibal and Johnson, 2015). Then, we tokenize it where each MWE is treated as a single token with an underline between the first and the second part (e.g. ***blow the whistle*** is mapped to ***blow_whistle***). Finally, we build three semantic spaces, using Word2Vec, Context2Vec, and BERT. Our goal is to determine the correct sense of an MWE in context, based on a manually tagged dataset, VNC (Cook et al., 2008). Following Melamud et al. (2016), we use the simple non-parametric version of the kNN classification algorithm (Cover and Hart, 1967) with k = 1 and for the distance measure, we rely on cosine distance of the vectors. As we do not do any extra training on the dataset, we divide it into evaluation and test sets. To classify a test MWE instance in context, we consider all the instances of the same MWE in the evaluation set and find the instance whose context embedding is the most similar to the context embedding of the test instance using the context-to-context similarity measure. Finally, we use the label of that instance as the correct label for the MWE in the test set. The rationale behind such a simple classification model is to make the comparison between the representations easy so that each model's success can be attributed directly to the input representations.

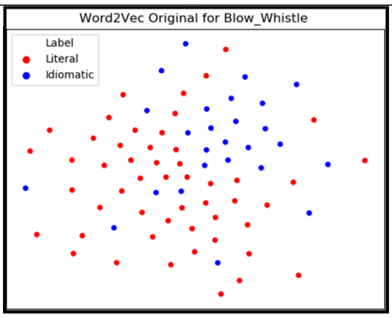
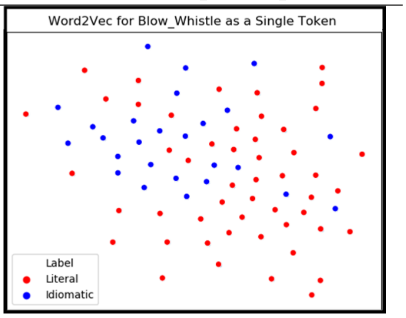
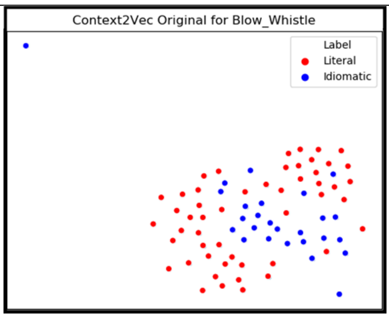
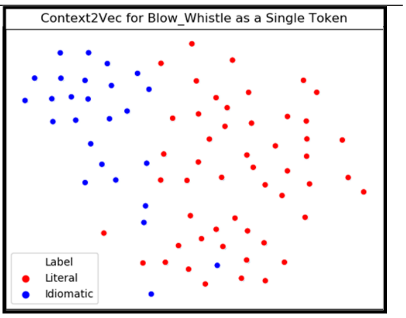
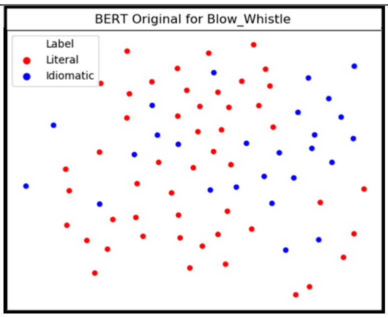
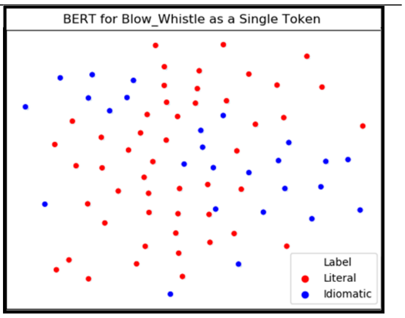
Model	Original	Idiom-Principle-inspired
Word2Vec		
Context2Vec		
BERT		

Table 3: t-SNE plots of different senses of 'blow the whistle' and their contextualized embeddings. The literal sense is in red and the idiomatic sense is in blue. Here, the VNC dataset is used.

5 Dataset and Evaluation

We use VNC-Tokens dataset (Cook et al., 2008) to evaluate our models. The dataset includes sentences containing Verb-Noun Combinations (VNC) tokens labelled as either idiomatic (I) / literal (L) (or "unknown"). For our experiments, we only use VNCs that are annotated as I or L as in King and Cook (2018). We evaluate the models using five-fold cross-validation and calculate the precision, recall, and F-score per each expression and then report the average scores as in King and Cook (2018), the results are reported per sense.

We also investigate to see how well different models encode information such as distinguishable senses in their vector space.

6 Experimental Results

In this section, we report the results of the first set of experiments where we create context representation using the original pre-trained models and then we present the results of the second set of experiments in which inspired by the Idiom Principle, we train our own models by treating each MWE as a single token. Then context embeddings are inferred using these trained models. Table 1 shows the results for the original pre-trained embeddings. As it can be seen the CWE, i.e. Context2Vec and BERT outperform the non-CWE, i.e. Word2Vec, up to 7% higher average F-score. In the next rows, Table 1 shows the results for Idiom Principle-inspired models along with those reported by King and Cook (2018). As it can be seen, the average F-score is the same for BERT and 1% higher for Word2Vec compared to the original models. However, both models achieved higher F-scores in detecting literal sense of MWEs. As for Context2Vec, the results improved by 6% on average and up to 10% in detecting literal sense of MWEs. We used an ANOVA test to check the statistical significance of the results of our models and found all our results to be significant at $p < 0.05$.

We did not expect the results to improve for Word2Vec as it always conflates the senses so it will not be able to learn different embeddings for different senses no matter how MWEs are treated in pre-processing step. In regard to BERT, we cannot see the improvement observed for Context2Vec. We speculate this might be due to the models inability to provide quality embeddings for rare words (Schick and Schütze, 2019) as treating each MWE as a single token turns it into a rare word for which the models need to learn an embedding. We will be investigating this in our future work. Nevertheless, the improvement on the results are noticeable (even for BERT) as we used a much smaller corpus to train our models compared to those used by the original models. We used BNC which contains 100 million words whereas the original models were trained on the corpora of much bigger size, namely Google News dataset with 100 billion words for Word2Vec, ukWaC (Ferraresi et al., 2008) with 2 billion words for Context2Vec, and the entire Wikipedia with about 2,500 million words and a book corpus with 800 million words for BERT.

King and Cook (2018) reported 0.69 F-score for the same dataset and then they used extra feature, the expression being in its canonical form or not, and increased the F-score to 0.77. However, this method is limiting as it requires feature engineering while our model of Context2Vec is capable of producing on par results, 0.75, without any external knowledge and by only relying on the features extracted by the model itself. We also used much smaller corpus to train our model in comparison with what they used, which was a snapshot of Wikipedia from September 2015, consisting of approximately 2.6 billion tokens (King and Cook, 2018). Moreover, they did extra training on the dataset after extracting the embeddings from their model whereas we did not do any training on the dataset.

To see how robust the models are across different expressions, we created the box plot for the models using the average F-scores of the models per expression. This is illustrated in Figure 1 and Figure 2 of Table 2 which shows the most robust model is BERT in the first setting and Context2Vec in the second setting. The robustness of a model is important as we do not want a model to work well for one MWE and poor for the other.

We are also curious to see whether the models are capable of placing different senses of an expression in different segments of their semantic space. For this, we use t-SNE (van der Maaten and Hinton, 2008) to map these high-dimensional spaces into two dimensional spaces. Table 3 shows t-SNE plots of two different senses of the expression ***blow the whistle*** in the VNC dataset encoded by the six different models in two settings: Original and Idiom-Principle-inspired.

As you can see, the Word2Vec and BERT embeddings hardly allow to distinguish any clusters as the senses are scattered across the entire plot, both in the original and Idiom-Principle-inspired settings. However, in Context2Vec embedding space, senses are placed in clearly separable clusters especially in the Idiom-Principle-inspired setting. This made us dig deeper into the Context2Vec model and probe its level of understanding through a lexical substitution task. In the lexical substitution task, the goal is to find a substitute word for a given target word in sentential context. To do so, we remove the MWE from a sentence and then get the embeddings of the remaining sentence which is in fact the context of the MWE. Then we find the embeddings of which words have the highest cosine similarity with the embeddings of the context. Table 4 shows the list of lexical substitutes proposed by the model for three MWEs per their literal/idiomatic senses. As you can see the model seems to be able to distinguish well between literal and idiomatic senses as it suggest suitable substitute for the removed MWE. The sentences are listed in Table 5.

MWE	Sense	Sentence #	Proposed Substitute
Kick heel	I	1	**wait**, stay, stop
	L	2	**Clap**, barefoot, **kick**
Hit road	I	3	go, **start, embark**
	L	4	**smash**, drop, shoot

Table 4: The lexical substitutes proposed by Context2Vec to replace MWEs in their literal or idiomatic senses.

Sentence #	Sentence
1	The man won't step foot outside his castle without myself as escort so I have to ***[kick my heels]*** until his business with Queen Matilda is done
2	I could see I was going to get warmer still because the bullock was beginning to enjoy the game ***[kicking up his heels]*** and frisking around after each attempt
3	The Ulster Society are about to ***[hit the road]*** on one of their magical history tours
4	The bullets were ***[hitting the road]*** and I could see them coming towards me a lot faster than I was able to reverse.

Table 5: List of sentences that are referred to in Table 4 by their number.

7 Conclusion and Future Work

In this work, we used Contextualized Word Embeddings (CWE), i.e. Context2Vec and BERT to include contextual information for distinguishing between the idiomatic and literal senses of an idiom in context. Moreover, inspired by the Idiom Principle, we hypothesized that to fully exploit the capacity of CWE, an idiom should be treated as a single token both in training and testing; The results showed that by applying Idiom Principle to CWE, especially Context2Vec, we can build a model to distinguish between literal and idiomatic senses of a potentially idiomatic expression in context. Through dimensionality reduction and lexical substitution, we further showed that Context2Vec is capable of placing literal and idiomatic senses in distinct regions of semantic space; Besides, the model has a good level of understating of the meaning as it suggests suitable replacement for both literal and idiomatic senses of set of MWEs. In our future work, we are interested in improving the results for BERT. We also would like to train the Idiom-Principle-inspired models on a bigger corpus to investigate how the results compare to what were achieved here.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Lou Burnard. 2000. The British National Corpus Users Reference Guide. Library Catalog: www.natcorp.ox.ac.uk.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions - MWE '07*, pages 41–48, Prague, Czech Republic. 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.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany. Association for Computational Linguistics.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations. page 9.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac , a very large web-derived corpus of english.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional Structure. *WORD*, 10(2-3):146–162.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Milton King and Paul Cook. 2018. https://doi.org/10.18653/v1/P18-2055 Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *ArXiv*, abs/1506.06726.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Jing Peng, Anna Feldman, and Hamza Jazmati. 2015. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 507–511.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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.

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.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2890–2895, Online. 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.

Timo Schick and Hinrich Schütze. 2019. Rare Words: A Major Problem for Contextualized Embeddings And How to Fix it by Attentive Mimicking. *arXiv:1904.06707 [cs]*. ArXiv: 1904.06707.

J. Sinclair, L. Sinclair, and R. Carter. 1991. *Corpus, Concordance, Collocation*. Describing English language. Oxford University Press.

Anna Siyanova-Chanturia and Ron Martinez. 2014. The Idiom Principle Revisited. *Applied Linguistics*, page amt054.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

A. Wray. 2002. https://books.google.sc/books?id=h2kpuAAACAAJ *Formulaic Language and the Lexicon*. Cambridge University Press.