Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space

Filip Klubička, Vasudevan Nedumpozhimana, John D. Kelleher
ADAPT Centre, Technological University Dublin, Ireland
filip.klubicka@adaptcentre.ie
{vasudevan.nedumpozhimana,john.d.kelleher}@TUDublin.ie

Abstract

The goal of this paper is to learn more about how idiomatic information is structurally encoded in embeddings, using a structural probing method. We repurpose an existing English verbal multi-word expression (MWE) dataset to suit the probing framework and perform a comparative probing study of static (GloVe) and contextual (BERT) embeddings. Our experiments indicate that both encode some idiomatic information to varying degrees, but yield conflicting evidence as to whether idiomaticity is encoded in the vector norm, leaving this an open question. We also identify some limitations of the used dataset and highlight important directions for future work in improving its suitability for a probing analysis.

1 Introduction

In recent years the NLP community has become somewhat enamoured by research on probing vector embeddings (Ettinger et al., 2016; Shi et al., 2016; Veldhoen et al., 2016; Adi et al., 2017) and justifiably so, as the method allows researchers to explore linguistic aspects of text encodings and has broad application potential. To date, however, the majority of impactful probing work focuses on analysing syntactic properties encoded in language representations, and the rich and complex field of semantics is comparably underrepresented (Belinkov and Glass, 2019; Rogers et al., 2020). One semantic problem that has received relatively little attention is the question of how models encode idiomatic meaning.

Laterally, our recently-developed extension of the probing method called *probing with noise* (Klubička and Kelleher, 2022) allows for structural insights into embeddings, highlighting the role of the vector norm in encoding linguistic information and showing that the norm of various embeddings can contain information on various surface-level, syntactic and contextual linguistic properties, as well as taxonomic ones (Klubička and Kelleher, 2023).

We hypothesise that probing idiomatic usage is a relevant task for studying the role of the norm: given there is some agreement that idiomatic phrases are at least partially defined by how strongly they are linked to the cohesive structure of the immediate discourse (Sag et al., 2002; Fazly et al., 2009; Sporleder and Li, 2009; Feldman and Peng, 2013; King and Cook, 2017), our intuition is that an idiomatic usage task should behave similarly to contextual incongruity tasks such as bigram shift and semantic odd-man-out (Conneau et al., 2018), which have been shown to be at least partially stored in BERT's vector norm (Klubička and Kelleher, 2022). For example, the idiomatic usage of a phrase such as *spill the beans* should have a similarly confounding effect on the sentence's word co-occurrence statistics as a semantic odd-man-out. This reasoning aligns with the findings of Nedumpozhimana and Kelleher (2021) who find that BERT can distinguish between sentence disruptions caused by missing words and the incongruity caused by idiomatic usage. Based on this, we are inclined to view an idiomatic usage task as a contextual incongruity task, and would expect to find some information stored in the norm.

To study this we repurpose an existing idiom token identification dataset into a probing task dataset and run it through our *probing with noise* pipeline, using both static GloVe and contextual BERT embeddings. Interestingly, while our experiments show that both GloVe and BERT generally do encode some idiomaticity information, the norm's role in this encoding is inconclusive, and further analysis points to some surprising irregularities in the behaviour of the models, which we trace back to a number of limitations in the dataset.

2 Related Work

Probing in NLP is defined by Conneau et al. (2018) as a classification problem that predicts linguistic properties using dense embeddings as training data.

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 45–57

The idea is to train a classifier over embeddings produced by some pretrained model, and assess the embedding model's knowledge encoding via the probe's performance. The framework rests on the assumption that the probe's success at a given task indicates that the encoder is storing information on the pertinent linguistic properties.

Given that embeddings are vectors positioned in a shared multidimensional vector space, we are interested in the structural properties of the linguistic information that they encode. Vectors are geometrically defined by two aspects: having both a **direction** and **magnitude**. Direction is the position in the space that the vector points towards (expressed by its dimension values), while magnitude is a vector's length, defined as its distance from the origin (expressed by the vector norm). Information contained in a vector is commonly understood to be encoded in the dimension values, however we have shown that it is also possible for the vector magnitude—the norm—to carry information as well (Klubička and Kelleher, 2022).

This is an important consideration for embedding research as it has been shown that normalising vectors removes information encoded in the norm (Goldberg, 2017; Klubička and Kelleher, 2022). A key step in calculating a cosine similarity measure, which is commonly used as a proxy for word similarity, is to normalise the vectors being compared. This has the side effect of nullifying any distinguishing properties the norms might have and any linguistic information encoded in the norm will be lost when making the comparison, which is an undesirable outcome if one wished to consider it in the comparison. We are thus interested in exploring how idiomaticity is encoded in vector space and whether any of it can be found in the norm.

The term Multi-Word Expression (MWE) frequently encompasses a wide variety of phenomena such as idioms, compound nouns, verb particle constructions, etc. The precise definition sometimes differs depending on the community of interest (Constant et al., 2017), and in this paper we use the terms *MWE*, *idiom* and *idiomatic phrase* somewhat liberally to mean any construction with idiomatic or idiosyncratic properties. This is sufficient for our interest in probing for a general notion of idiomaticity, the difference between idiomatic and literal usage of MWEs and studying how this distinction is encoded by embedding models.

Notably, as probing is a relatively recent framework and idioms are still a difficult phenomenon to model, not much work has been done in this space. Some inspiration can be found in the idiom token identification literature, closely related to word-sense disambiguation, where the goal is to build models that can discriminate idiomatic from literal usage (Hashimoto and Kawahara, 2008, 2009; Fazly et al., 2009; Li and Sporleder, 2010a,b; Peng et al., 2014; Salton et al., 2017; Peng and Feldman, 2017; King and Cook, 2018; Shwartz and Dagan, 2019; Hashempour and Villavicencio, 2020). While they do not overtly apply probing in their work, Salton et al. (2016) were the first to use an idiom token identification pipeline that is comparable to a typical probing framework, where sentence embeddings are used as input to a binary classifier that predicts whether the sentence contains a literal or figurative use of a MWE, indicating that an idiom probing task can be successful.

We have built upon this notion and performed sentence-level probing for idiomaticity in BERT (Nedumpozhimana et al., 2022). We employed the game theory concept of Shapley Values (Shapley, 1953) to rank the usefulness of individual idiomatic expressions for model training, in an effort to identify the types of signal that BERT captures when modelling idiomaticity. This approach has revealed that providing training data that maximises coverage across topics is the most useful form of topic information, and our findings indicate that there is no one dominant property that makes an expression useful, but rather fixedness and topic features are combined contributing factors. This current paper presents a successor study, as we now look for structural traces of idiomaticity at the sentence level. However, recently there have also been some interesting word-level probing studies.

Nedumpozhimana and Kelleher (2021) perform word-level probing experiments on BERT, where they combine probing with input masking to analyse the source of idiomatic information in a sentence, and what form it takes. Results indicate that BERT's idiomatic key is primarily found within an idiomatic expression, but also draws on information from the surrounding context. Meanwhile, Garcia et al. (2021) use probing to assess if some of the expected linguistic properties of idiomatic noun compounds and their dependence on context and sensitivity to lexical choice can be extracted from contextual embeddings. They conclude that idiomaticity is not yet accurately represented by contextual models: while they might be able to detect idiomatic usage, they may not detect that idiomatic noun compounds have a lower degree of substitutability of their individual components.

When it comes to idiomatic probing benchmarks, the Noun Compound Senses Dataset (Garcia et al., 2021) is the only curated idiomaticity probing dataset. Other idiom probing work (Salton et al., 2016; Nedumpozhimana and Kelleher, 2021; Nedumpozhimana et al., 2022) relies on existing MWE and idiom datasets, specifically the VNC-tokens dataset (Cook et al., 2008). Other MWE resources for English include the PARSEME working group's (Savary et al., 2017; Ramisch et al., 2018) VMWE dataset, (Walsh et al., 2018), the STREUSLE corpus (Schneider and Smith, 2015) and a verbal MWE dataset by Kato et al. (2018). However, these are annotated at the word-level, employ a fine-grained taxonomy of labels and only annotate idiomatic usage of MWEs, making it impossible to train models that can differentiate between literal and idiomatic usage. As such, while meticulously crafted and, as we argue in §7.1, of much higher quality than what we use in our work, they are not suited for the type of sentence-level analysis of idiomaticity we are interested in. There are recent datasets that are better suited for this: MAGPIE (Haagsma et al., 2020) and the SemEval-2022 Task 2 dataset (Tayyar Madabushi et al., 2022). Unfortunately we only became aware of the former during the review process, while the latter was not yet freely available at the conception of this research. Instead, to stay consistent with the recent wave of idiom probing work, we repurpose the the VNC-tokens dataset (Cook et al., 2008) to suit our structural probing needs, as presented in §3.

3 Probing Dataset Construction

Our *Idiomatic Usage* (IU) task is based on the VNC-Tokens dataset (Cook et al., 2008), which is a collection of English sentences containing MWEs called Verb-Noun Combinations (VNC), which can be used idiomatically or literally. This includes expressions such as *hit road*, *blow whistle*, *make scene* and *make mark*. The VNC-tokens dataset contains a total of 2,984 sentences with 56 different expressions, with each sentence containing one expression. Each sentence in the dataset is labelled as *Idiomatic*, *Literal*, or *Unknown*. However, the related literature only makes use of a subset of the full dataset. For consistency and comparability with related work (Peng et al., 2014; Salton et al., 2016; Nedumpozhimana and Kelleher, 2021) we apply the same filtering heuristics so the subset used in our experiments contains 1,205 sentences, of which 749 are labelled as *Idiomatic* and 456 are labeled as *Literal*, allowing for straightforward binary classification. A breakdown of each expression in the dataset is displayed in Table 7 in Appendix A.

verb	noun
make	face, pile, hay, scene, mark, hit
pull	leg, weight, plug, punch
blow	whistle, top, trumpet
hit	wall, roof, road
get	wind, sack, nod
lose	head, thread

Table 1: Groups of VNCs based on verb constituent overlap.

3.1 Choosing the right train and test split

In establishing a train and test split we aimed to avoid lexical memorisation (Levy et al., 2015; Santus et al., 2016; Shwartz et al., 2017), as our goal is for the probe to only learn a general, abstract, notion of idiomaticity unrelated to any particular idiomatic phrase, so the train and test sets need to be carefully curated. We tackle this on two fronts:

(a) The probe needs to be tested on a subset of VNCs that it has not seen in training. Having it predict the usage status of only unfamiliar idiomatic phrases forces the model to fall back on its general knowledge of what makes an idiomatic phrase, rather than a memory of any specific VNC.

(b) When training, we also need to ensure that the model attends to general properties of idiomaticity, rather than phrase- or token-specific ones. The surface form of a VNC likely has significant informational value to either the encoder or the probe, so specific VNC constituents might be interpreted as some sort of signal. Upon inspection of the candidate phrases we have found that many of the 28 VNCs in the dataset share the same verb constituent, as shown in Table 1. In fact, the dataset contains only 7 VNCs that contain "unique" verb constituents: *hold fire*, *have word*, *take heart*, *kick heel*, *see star*, *cut figure*, *find foot*.

We attempt to mitigate this by populating the train set exclusively with phrases with overlapping verbs, while placing the phrases with unique verbs in the test set. Thus the importance of individual verbs is reduced as they appear with different nouns. Coincidentally, satisfying condition (b) also satisfies condition (a), so no additional filtering is needed: VNCs from the test set do not appear in the train set, and the usage of verbs in the train set is diverse with different VNCs having the same verb constituent. As such, our test set includes 7 VNCs, while the remaining 21 are used in training. Table 8 in Appendix A shows the final train and test split used in our experiments.

Additionally, to confirm that the chosen train and test split is a viable way to tease out idiomaticity, we also run a parallel set of experiments using a form of bootstrapping where we resample the train and test split multiple times by randomly choosing 7 VNCs to be used in the test set, and using the remaining 21 phrases for training. This violates the above-established principle (b) as verbal constituents might be mixed between train and test sets, but still conforms to principle (a), as the model will always be tested on a set of 7 phrases that were not seen during training. Additionally, as we are not fixing the number of samples in the train and test sets, but rather the number of idiomatic phrases (with a varying number of sentences containing each phrase), there will also be slight differences in the ratio of the train and test sample sizes between different runs. However, we find that when the multitude of runs are averaged the true effect comes to the fore—the bootstrapped results mirror the results of the fixed setting, confirming the chosen split. For transparency and completeness, in Section 5 we report results for both setups: Idiomatic Usage Fixed data split (IU_F) and Idiomatic Usage Resampled data split (IU_R).

4 Experimental Setup

4.1 Chosen Embeddings

Given the prominence of contextual encoders such as BERT (Devlin et al., 2019) and its derivatives, as well as their ability to model in-context meaning and incongruity, they are an obvious choice for our analysis. However, rather than compare different contextual encoders, we prefer to draw a contrastive comparison with a static encoder such as GloVe (Pennington et al., 2014), which is based on a word to word co-occurrence matrix, as this comparison can provide more varied insight.

Given that our idiomatic usage dataset is framed as a classification task at the sentence level, our experiments require sentence representations. We use pretrained versions of BERT and GloVe to generate embeddings for each sentence. The BERT model generates 12 layers of embedding vectors with each layer containing a separate 768-dimensional embedding for each word, so we average the word embeddings in BERT's final layer, resulting in a 768-dimensional sentence embedding. We take the same mean pooling approach with GloVe, which yields a 300-dimensional sentence embedding for each sentence. While BERT uses sub-word tokens to get around out of vocabulary tokens, in the rare instance of encountering an OOV with GloVe, we generate a random word embedding in its stead.

4.2 Probing with Noise

The method is described in detail in Klubička and Kelleher (2022)[1]: in essence it applies targeted noise functions to embeddings that have an ablational effect and remove information encoded either in the norm or dimensions of a vector.

We remove information from the norm (abl.N) by sampling random norm values and scaling the vector dimensions to the new norm. Specifically, we sample the L2 norms uniformly from a range between the minimum and maximum L2 norm values of the respective embeddings in our dataset.[2]

To ablate information encoded in the dimensions (abl.D), we randomly sample dimension values and then scale them to match the original norm of the vector. Specifically, we sample dimension values uniformly from a range between the minimum and maximum dimension values of the respective embeddings in our dataset.[3] We expect this to fully remove all interpretable information encoded in the dimension values, making the norm the only information container available to the probe.

Applying both noise functions to the same vector (abl.D+N) should remove any information encoded in it, meaning the probe has no signal to learn from, a scenario equal to training on random vectors.

Even when an embedding encodes no information, our train set contains class imbalance and the probe can learn the distribution of classes. To account for this, as well as the possibility of a powerful probe detecting an empty signal (Zhang and Bowman, 2018), we establish informative random

[1] Code available here: https://github.com/GreenParachute/probing-with-noise

[2] GloVe: [2.2634,4.2526]
BERT: [7.4844,11.1366]

[3] GloVe: [-1.7866, 2.8668]
BERT: [-5.0826, 1.5604]

baselines against which we compare the probe's performance. We employ two such baselines: (a) we assert a random prediction (*rand.pred*) onto the test set, negating any information that a classifier could have learned, class distributions included; and (b) we train the probe on randomly generated vectors (*rand.vec*), establishing a baseline with access only to class distributions.

Finally, to address the degrees of randomness in the method, we train and evaluate each model 50 times and report the average score of all the runs, essentially bootstrapping over the random seeds (Wendlandt et al., 2018). Additionally, we calculate a confidence interval (CI) to ensure that the reported averages were not obtained by chance, and report it alongside the results to indicate statistical significance when comparing averages.

4.3 Probing Classifier and Evaluation Metric

In our experiments the sentence embeddings are used as input to a Multi-Layered Perceptron (MLP) classifier, which labels them as idiomatic (1) or literal (0). We evaluate the performance of the probe using the micro-average AUC-ROC score,[4] the most appropriate evaluation metric for a dataset with unbalanced labels, as it reflects the classifier's performance on both positive and negative classes. Regarding implementation and parameter details, we used the bert-base-uncased BERT model from the *pytorch_pretrained_bert* library[5] (Paszke et al., 2019), a pre-trained GloVe model[6] and for the MLP probe we used the scikit-learn MLP implementation (Pedregosa et al., 2011) using the default parameters.[7]

5 Experimental Results

Experimental evaluation results for GloVe and BERT on the idiomatic usage (IU) probing task are presented in Tables 2 and 3. The tables include results for both the setting where the VNC's in the hold-out test set are fixed (IU_F) and the setting where they are resampled each time (IU_R), though this is essentially the same probing task. Note that cells shaded light grey belong to the same distribution as random baselines, as there is no statistically significant difference between the different scores; cells shaded dark grey belong to the same distribution as the vanilla baseline; and cells that are not shaded contain a significantly different score than both the random and vanilla baselines, indicating that they belong to different distributions.

GloVe				
Model	$\mathbf{IU_F}$		$\mathbf{IU_R}$	
	auc	±CI	auc	±CI
rand. pred.	.4994	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5	.0013
vanilla	.7485	.0003	.7717	.0022
abl. N	.7445	.0006	.7687	.0021
abl. D	.5012	.0018	.4993	.0015
abl. D+N	.4991	.0018	.5005	.0015

Table 2: Probing results on GloVe models and baselines, both with fixed (F) and resampled (R) test set. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

BERT				
Model	$\mathbf{IU_F}$		$\mathbf{IU_R}$	
	auc	±CI	auc	±CI
rand. pred.	.4997	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5013	.0013
vanilla	.8411	.0002	.8524	.0016
abl. N	.8413	.0003	.8532	.0016
abl. D	.4991	.0019	.4978	.0015
abl. D+N	.4999	.0018	.5004	.0015

Table 3: Probing results on BERT models and baselines, both with fixed (F) and resampled (R) test set. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

The results interpretation here is quite straightforward. As the unablated, vanilla baseline significantly outperforms random baselines in both models, this indicates that both GloVe and BERT encode a non-zero amount of idiomatic usage information, which aligns with previous findings.

$\mathbf{IU_F}$ vs. $\mathbf{IU_R}$: It important to validate our chosen train and test split (see §3.1) by comparing the respective vanilla performances of IU_F and IU_R. Given that our goal is to nudge the probe to model a representation of idiomaticity that is unrelated to any given phrase, we expect that the IU_F setting

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

[5] https://pypi.org/project/pytorch-pretrained-bert/

[6] The larger common crawl vectors: https://nlp.stanford.edu/projects/glove/

[7] activation='relu', solver='adam', max_iter=200, hidden_layer_sizes=100, learning_rate_init=0.001, batch_size=min(200,n_samples), early_stopping=False, weight init. $W \sim \mathcal{N}\left(0, \sqrt{6/(fan_{in} + fan_{out})}\right)$ (scikit relu default). See: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

should make the task more difficult for the classifier. The results confirm this, showing that in GloVe and BERT vanilla IU_R significantly outperforms vanilla IU_F. Evidently, the curated test split makes prediction on the task more challenging and the lower performance of IU_F indicates that the model is forced to rely on VNC-independent features to make predictions.

GloVe vs. BERT: In terms of differences between encoders, the results show that vanilla BERT significantly outperforms vanilla GloVe in both the IU_F and IU_R scenarios. Evidently, BERT is much better at encoding idiomaticity than GloVe. We suspect this is due to two factors: (a) BERT is a contextual encoder and as such is better suited to modelling the local context necessary to accurately represent idiomaticity in the sentence, and (b) it has a much higher dimensionality, meaning it has the potential to devote more representation space to more complex phenomena.

Idiomaticity and the norm: One of the goals of this experiment was to investigate whether the norm encodes any information relevant to the IU task. Our method states this is most clearly determined in the setting with ablated dimension information (abl.D), where above random performance indicates that the information is stored in the unablated norm container (Klubička and Kelleher, 2022). Our results here show no conclusive indication that the norm encodes idiomaticity information on this task: in all four scenarios ablating only the dimensions already makes the probe's performance comparable to random, which indicates no information is stored in the norm.[8]

As stated in the introduction, given the IU task's similarity with contextual incongruity tasks, we would expect to find some signal in the norm. Our result here is somewhat surprising and motivates further questions, prompting us to perform additional post hoc investigations and analyses that should improve our understanding of the results and help shape our overall findings.

6 Additional Experiments

6.1 Norm Correlation Analysis

For another perspective on the relationship between vector norms and the IU task information, we run a post hoc analysis on the norm container. We investigate both the norms of our embeddings using a Pearson correlation analysis, which can be considered a linear probing study: we test the correlation between each vector norm (L1 and L2) and the sentence labels (*Idiomatic* and *Literal*[9]). The correlation results are presented in Table 4 and seem to be somewhat at odds with our experimental results.

Task	Vectors	GloVe		BERT	
		L1	L2	L1	L2
IU	vanilla	-0.2231	-0.1786	-0.1490	-0.1756
	abl. N	-0.0074	0.0276	-0.0397	-0.0167

Table 4: Pearson correlation coefficients between class labels and L1 and L2 norms for vanilla vectors and vectors with ablated norms. For this analysis the Idiomatic label was mapped to 1 and the Literal label to 0.

The analysis shows that in both vanilla GloVe and BERT both norms have a weak negative correlation with IU labels. While the correlations are weak, they are not zero—we observe a significant drop in the coefficients upon applying the norm ablation function, which seems to fully remove information from both norms, as the correlation coefficients drop to ≈0, indicating that relevant information encoded in the norms has been removed.

This difference between vanilla and abl.N points to some slight correlation between the idiomaticity labels and information encoded in the vanilla norm, yet our probing experiments do not align with this finding. What makes this more unusual is that our IU correlations are comparable to the correlations on parse tree depth (0.1908) or semantic odd-man-out (0.2305) tasks which do produce a signal in the *probing with noise* experiments as previously reported Klubička and Kelleher (2022).

It is possible that the correlation is just on the verge of being too weak to be detectable by the method. On the other hand, this could be a sign that other factors are at play—we suspect that the misalignment between the probing and correlation results hints at the imbalanced nature of the IU dataset and its limitations. We run an additional experiment to search for more evidence.

As an aside, it is worth noting that if we were to take the correlation results at face value, they do provide some interesting insight into how idiomatic usage is encoded in vector space. Specifically, a non-zero negative correlation coefficient

[8] We do see a hint of this when ablating the norm in GloVe IU_F, but this is more likely a feature of this particular data split, as the signal is not mirrored in IU_R. Even if it was, without a signal in the abl.D setting, the abl.N setting is insufficient evidence to infer that the norm encodes information.

[9] The Pearson test only works on continuous variables, but it is still possible to calculate with categorical variables if they are binary, by simply converting the categories to 0 and 1.

GloVe				
Model	**IU$_F$**		**IU$_R$**	
	auc	±CI	auc	±CI
rand. pred.	.4994	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5	.0013
vanilla	.7485	.0003	.7717	.0022
del. 1h	**.7737**	.0005	.7553	.0023
del. 2h	.7043	.0005	.7545	.002

Table 5: Probing results on GloVe dimension deletions both with fixed (F) and randomised (R) test set. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

BERT				
Model	**IU$_F$**		**IU$_R$**	
	auc	±CI	auc	±CI
rand. pred.	.4997	.0015	.4998	.0013
rand. vec.	.4997	.0015	.5013	.0013
vanilla	.8411	.0002	.8524	.0016
del. 1h	**.8668**	.0002	**.8576**	.0016
del. 2h	.8137	.0003	.8368	.0016

Table 6: Probing results on BERT dimension deletions both with fixed (F) and randomised (R) test set. Reporting average AUC-ROC scores and confidence intervals (CI) of the average of all training runs.

means that sentences containing idiomatic usage are positioned closer to the origin relative to sentences that contain literal usage. In other words, both GloVe and BERT vectors of sentences containing idiomatic usage are slightly shorter, which is an intriguing structural finding.

6.2 Dimension Deletion

We run supplementary experiments to investigate the role of the dimension container as the sole carrier of IU information. To do this we perform a dimension deletion experiment. Partially inspired by the work of Torroba Hennigen et al. (2020) who found that most linguistic properties are reliably encoded by only a handful of dimensions, we attempt to roughly identify the degree of localisation of information in the vector dimensions. In staying consistent with the ablational nature of the method, we simply delete one half of the vector's dimensions and retrain the probe on the truncated vectors, repeating the process for the remaining half.

The dimension deletion results are included in Tables 5 and 6. In these tables the row denoted *del.1h* reports the results for deleting the 1st half of an embedding vector, and *del.2h* reports results for deleting the 2nd half. Given that all relevant IU information seems to be encoded in vector dimensions, we expect that deleting half of the vector would cause a significant drop in performance when compared to vanilla. We would also expect a drop in evaluation scores regardless of which half of the vector is deleted. However, our results reveal some rather surprising effects.

While *del.2h* in GloVe causes the expected performance drop, in IU$_F$ *del.1h* causes a statistically significant *improvement* when compared to the vanilla baseline (marked in bold). We observe quite a large performance spike, though this is not mirrored in the IU$_R$ scenario. We might dismiss this as just a strange artefact of the particular IU$_F$ data split, were it not for the fact that we observe the same behaviour in both IU$_F$ and IU$_R$ in BERT, where *del.2h* causes a significant performance drop, but *del.1h* causes a significant spike.

It seems that both GloVe and BERT exhibit a certain degree of information localisation, with a preference for storing relevant IU information in the first half of dimensions, to the point where the second half reduces the overall information quality of the vector. In principle this interpretation is consistent with the findings of Torroba Hennigen et al. (2020) and Durrani et al. (2020), who showed that certain linguistic properties are localised in dimensions of contextual embeddings. However, we remain skeptical and wonder whether our findings reflect how these embeddings truly encode idiomaticity, or whether this is property of this particular dataset. We consider this in the following section.

7 Discussion and Limitations

While the correlation coefficients between both GloVe's and BERT's norm and the IU labels are non-zero, our probe does not seem to be able to leverage this information from the norm. In isolation, the correlation coefficient would have led us to believe that there may be some idiomaticity information encoded in the norm. However, this has not been confirmed by the *probing with noise* method, which when used in conjunction with the correlation analysis offers conflicting evidence.

The performance spikes exhibited in the deletion experiments are somewhat baffling, especially given the stark differences between the GloVe and BERT architectures. However, if the IU task were truly analogous to a contextual incongruity task, then arguably vanilla GloVe should be much worse at encoding IU than shown in our results—by design, an averaged GloVe sentence embedding cannot be aware of word order or relationships between words in a specific context and should perform much more poorly on such tasks, making even vanilla GloVe's performance a result that raises more questions than it answers.

One pertinent consideration regards the fact that our experiments were performed at the sentence level. It is possible that there is a crisper signal in the norm of individual word embeddings (as shown on a word-level taxonomic probing task (Klubička and Kelleher, 2023)). Averaging word embeddings to obtain sentence representations may have diluted the signal to the point where it is not detectable by the *probing with noise* method. Replicating our experiments at the word-level, or using more direct sentence representation approaches (such as using BERT's CLS token, doc2vec (Le and Mikolov, 2014) or SentenceBERT (Reimers and Gurevych, 2019)) might produce a more salient result.

As it stands, the majority of the results we have observed on the IU dataset behave like surprising outliers that are difficult to explain. This can either be due to strong confounding factors at play that we are not aware of or, perhaps more likely, this is evidence of our suspicion that the dataset is not well-suited for this type of analysis. And while we have learned that vanilla BERT is better at the task than GloVe, the question whether idiomaticity can be encoded in their norms remains an open one.

7.1 Dataset Limitations

While constructing and experimenting with the VNC-tokens dataset we have become aware of some of its shortcomings. Our main concern is that it is two orders of magnitude smaller than more established probing datasets (Conneau et al., 2018). While we addressed this by increasing the number of training runs and resampling the train and test set, its size still limits what the models are able to learn. Unfortunately, in dealing with an intricate phenomenon such as idioms, considerably-sized corpora are few and far between.[10]

The VNC-tokens dataset is also very limited in scope, containing only a single type of verbal MWE, while other datasets include a wider variety of verbal expressions or compounds involving other parts of speech. It is also worth noting that both idiomatic and literal usages of the VMWEs present in the dataset are relatively frequent in English when compared to other more niche idiomatic phrases. This relative frequency is likely also reflected in the pretrained embeddings and could affect a model's ability to model their idiomaticity, raising the question whether relatively rarer phrases might behave differently. Thus the generalisability of our findings to other idiomatic expressions is uncertain.

Furthermore, at this point the VNC-Tokens dataset is a relatively older benchmark and there are indications that it has not been as meticulously crafted as more recent MWE datasets. For example, the dataset does not control for sentence length, which could be a strong confounding factor, it contains some typographical errors, even some seemingly incorrect IU annotations, as well as literary language which contains OOV tokens for the pretrained GloVe model. It is our impression that cleaning up the dataset, aligning it with the PARSEME annotation guidelines[11], and updating it with additional examples of sentences containing VNCs in order to better balance the idiomaticity labels would greatly improve its overall quality.

Overall, in spite of our best efforts at mitigating confounders and constructing the right data split for our task, we still wonder whether the dataset is simply too small and too imbalanced to truly be useful in our probing scenario. Given all the limitations we have become aware of over the course of our experimentation it is difficult to decide whether our results are inconclusive due to the dataset, the type of idioms studied, perhaps some unknown limitation of the approach, or are simply a true observation. This makes our partially inconclusive and partially surprising findings somewhat difficult to reconcile with previous work. We thus emphasise the importance of expanding this work to a wider category of idiomatic phrases and ideally folding in all the datasets mentioned in §2—applying *probing with noise* to the datasets individually as well as an amalgamation of datasets would provide a more comprehensive analysis of general idiomaticity encoding and could provide more salient insights. It might also be beneficial to consider other dimensions of idiomaticity in the experimentation

[10] In fatct, all existing MWE resources are within a comparable size range to the VNC-tokens dataset. Even concatenating them would not nearly approach the size of probing datasets for non-semantic tasks.

[11] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=010_Definitions_and_scope/020_Verbal_multiword_expressions

and analysis, such as evaluating MWEs that are differentiated with respect to whether or not they carry a metaphorical mapping to literal usages, and whether or not they are grammatical or extragrammatical (Fillmore et al., 1988).

8 Conclusion and Future Work

In this paper we applied the *probing with noise* method to two different types of word representations—static and contextual—generated by two different embedding algorithms—GloVe and BERT—on a repurposed idiomatic usage probing task, with the aim of obtaining structural insights into the role of the norm encoding idiomatic usage information.

Overall we detect some mixed signals in our findings, which include that **(a)** generally both GloVe and BERT encode idiomatic usage information, but BERT encodes more **(b)** the norm of GloVe and BERT carries no idiomaticity information (or at least this is not recoverable by the probe), even though **(c)** it seems there is a correlation between the norm length and idiomatic usage in a sentence, where sentences containing idiomatic usage are positioned relatively closer to the origin of the vector space. **(d)** Additionally, it seems both GloVe and BERT prefer to store idiomatic usage information in the first half of their vectors, to the point where the second half is detrimental to the vector's overall encoding of idiomaticity. Finally, **(e)** we present these findings with the caveat that they only apply to the VNC-Tokens dataset, which requires a bit of a rework in order to be up to the standard required for a probing framework.

As for our initial research question, we asked whether embeddings models such as BERT might see an idiomatic usage task as being of the same category as a contextual incongruity task.[12] Given that vanilla BERT strongly outperforms vanilla GloVe on the task, this could lend some credence to the interpretation that contextual awareness and the ability to model incongruity, which GloVe lacks but BERT excels at, is what improves its idiomaticity encoding. However, evidence is inconclusive and whether the vector norm of either model plays a role in encoding idiomatic information in the same way that it supplements the encoding of contextual incongruity information remains an open question, which we are committed to further pursue in future work. This would involve cleaning the VNC-Tokens dataset and combining it with other existing MWE datasets in a systematic exploration of the structural encoding of idiomaticity.

[12]This hypothesis inspired the title of the paper, referring to Lakoff (1987) and his work on semantic categories.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreements No. 13/RC/2106 and 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme, and is co-funded under the European Regional Development Fund.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR, 2017*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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.

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

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.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. 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.

Afsanesh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. In *Computational Linguistics*, volume 35, pages 61–103.

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.

Charles J Fillmore, Paul Kay, and Mary Catherine O'connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.

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

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):117.

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.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on wsd incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 992–1001.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43(4):355–384.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Construction of large-scale English verbal multiword expression annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Milton King and Paul Cook. 2017. Supervised and unsupervised approaches to measuring usage similarity. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 47–52, Valencia, Spain. Association for Computational Linguistics.

Milton King and Paul Cook. 2018. 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.

Filip Klubička and John D Kelleher. 2023. Probing taxonomic and thematic embeddings for taxonomic information. In *Proceedings of the 12th International Global Wordnet Conference*.

Filip Klubička and John Kelleher. 2022. Probing with noise: Unpicking the warp and weft of embeddings. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 404–417, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

George Lakoff. 1987. *Women, fire, and dangerous things: What Categories Reveal about the Mind*, volume 10. University of Chicago press Chicago.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

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.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Vasudevan Nedumpozhimana, Filip Klubička, and John D. Kelleher. 2022. Shapley idioms: Analysing bert sentence embeddings for general idiom token identification. *Frontiers in Artificial Intelligence*, 5.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jing Peng and Anna Feldman. 2017. *Automatic Idiom Recognition with Word Embeddings*, pages 17–29. Springer International Publishing, Cham.

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.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ivan A. Sag, Thimothy 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: Third International Conference: CICLing-2002, Lecture Notes in Computer Science*, volume 2276, pages 1–15.

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.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine features in a random forest to learn taxonomical semantic relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4557–4564, Portorož, Slovenia. European Language Resources Association (ELRA).

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

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.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75, Valencia, Spain. Association for Computational Linguistics.

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*, pages 754–762.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Sara Veldhoen, Dieuwke Hupkes, and Willem H Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (at NIPS)*.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an annotated corpus of verbal MWEs for English. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193–200, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Expression	#samples	#idiomatic	ratio
see star	61	5	0.08
hit wall	63	7	0.11
pull leg	51	11	0.22
hold fire	23	7	0.30
make pile	25	8	0.32
blow whistle	78	27	0.35
make hit	14	5	0.36
get wind	28	13	0.46
lose head	40	21	0.53
make hay	17	9	0.53
make scene	50	30	0.60
hit roof	18	11	0.61
blow trumpet	29	19	0.66
make face	41	27	0.66
pull plug	64	44	0.69
take heart	81	61	0.75
hit road	32	25	0.78
kick heel	39	31	0.79
pull punch	22	18	0.82
pull weight	33	27	0.82
blow top	28	23	0.82
cut figure	43	36	0.84
make mark	85	72	0.85
get sack	50	43	0.86
have word	91	80	0.88
get nod	26	23	0.88
lose thread	20	18	0.90
find foot	53	48	0.91
TOTAL:	1205	749	0.62

Table 7: VNCs ordered by % of idiomatic usage: number of samples (#samples), number of idiomatic uses (#idiomatic) % of idiomatic usage (ratio).

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. 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 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

VNC	Train set Total	Idiomatic	VNC	Test set Total	Idiomatic
blow top	28	23			
blow trumpet	29	19			
blow whistle	78	27			
get sack	50	43			
get nod	26	23			
get wind	28	13			
hit road	32	25			
hit roof	18	11	cut figure	43	36
hit wall	63	7	find foot	53	48
lose head	40	21	have word	91	80
lose thread	20	18	hold fire	23	7
make face	41	27	kick heel	39	31
make hay	17	9	see star	61	5
make hit	14	5	take heart	81	61
make mark	85	72			
make pile	25	8			
make scene	50	30			
pull leg	51	11			
pull plug	64	44			
pull punch	22	18			
pull weight	33	27			
Total:	814	481		391	268
Ratio:		0.5909			0.6854

Table 8: A breakdown of VNCs and idiomatic instances in the train and test split.

A Appendix A

A.1 Dataset Statistics

In Table 7 the VNC expressions are listed by increasing order of percentage of idiomatic usage: *see star* is the expression with the lowest percentage of idiomatic usage (8.20%) and *find foot* is the expression with the highest percentage of idiomatic usage (90.57%). The overall percentage of idiomatic instances (regardless of the expression) is 62%.

Table 8 displays the final train and test split we used in our experiments, as well as a breakdown of specific expressions and their labels in both sets, sorted according to the verbal constituent. While this split is not focused on the ratio of training instances, but rather subsets of training instances containing the same VNC, this does mirror the 25%/75% data split employed by (Salton et al., 2016). Though the 68% ratio of idiomatic phrases in the test set is somewhat higher than maintained in previous work (≈62%), we expect the specific choices of VNCs will have a positive effect overall in priming the classifier to use its knowledge of idiomaticity to make predictions.