Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings

Murathan Kurfalı, Robert Östling
Linguistics Department, Stockholm University
Stockholm, Sweden
{murathan.kurfali,robert}@ling.su.se

Abstract

The majority of multiword expressions can be interpreted as figuratively or literally in different contexts which pose challenges in a number of downstream tasks. Most previous work deals with this ambiguity following the observation that MWEs with different usages occur in distinctly different contexts. Following this insight, we explore the usefulness of contextual embeddings by means of both supervised and unsupervised classification. The results show that in the supervised setting, the state-of-the-art can be substantially improved for all expressions in the experiments. The unsupervised classification, similarly, yields very impressive results, comparing favorably to the supervised classifier for the majority of the expressions. We also show that multilingual contextual embeddings can also be employed for this task without leading to any significant loss in performance; hence, the proposed methodology has the potential to be extended to a number of languages.

1 Introduction

By definition, a multiword expression (MWE) is idiomatic in the sense that its meaning cannot be derived from the meanings of its components. However, whereas sometimes a sequence of words corresponding to an MWE only has the idiomatic interpretation (e.g., *by and large*), there is often also a literal interpretation of the same sequence, resulting in an ambiguity:

- And the final twenty minutes is a headlong adrenalin rush, frantically intercutting four separate battle sequences and never **dropping the ball** once.
- Now, **drop the ball** for a bounce, tap it softly up towards your hands but let it fall back to the pavement for another bounce. *(examples taken from Korkontzelos et al. (2013))*

Such multiword expressions are commonly referred as *potentially idiomatic expressions (henceforth, PIE)* and determining the correct meaning of a PIE in context is shown to be crucial for many downstream tasks including sentiment analysis (Williams et al., 2015), automatic spelling correction (Horbach et al., 2016) and machine translation (Isabelle et al., 2017). Most of the previous work capitalizes on the differences between the contexts where PIEs are used idiomatically and literally. Following that insight, we investigate the applicability of recent contextual embedding models to disambiguation of PIEs.

Contextual embeddings, e.g. ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), have emerged in the last few years and quickly become the standard in a variety of tasks. These are very deep neural language models which are pre-trained on large-scale corpora. Unlike the conventional static word embeddings, such as Word2Vec (Mikolov et al., 2013) where each word type is represented by a fixed vector, these models assign distinct representations for each input token dependent on their context. Hence, they are called *contextual word embeddings*, highlighting their sensitivity to the context. For example, in the sentence "Can you throw this can away?" the first and second occurrence of the token *can* are supposed to be assigned substantially different embeddings.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 85–94
Barcelona, Spain (Online), December 13, 2020.

The extent of the contextuality of these embeddings, on the other hand, is still an open research topic (Ethayarajh, 2019). In this work, we specifically investigate whether such contextual embeddings provide sufficient contextual information to distinguish literal usages of PIEs from idiomatic ones. To this end, we represent the PIE tokens in a certain context by their corresponding BERT embeddings (Devlin et al., 2019) and perform both supervised and unsupervised PIE disambiguation. The results suggest that the plain BERT model, without any fine-tuning or further training, is able to encode the different usages of PIEs to the extent that, even a with simple linear classifier, we can substantially improve the state-of-the-art on common datasets in two different languages.

The unsupervised classification, on the other hand, is performed via hierarchical clustering, accompanied with a simple heuristic, that PIEs with literal interpretations are semantically closer to their context than the idiomatic ones. For the most of the time, the unsupervised classification also achieves unprecedented performance although not as consistently as its supervised counterpart, failing completely with some expressions.

Finally, we compare the performance of the monolingual BERT models with the multilingual-BERT (mBERT) to investigate the applicability of our approach to other low resource languages as well as to provide further insight regarding the cross-lingual capabilities of the multilingual contextual embeddings when they are employed directly; that is, without any fine-tuning in the target language. The results show that multilingual-BERT achieves comparable results to monolingual models across all datasets, suggesting that the proposed methodology can straightforwardly be extended to other languages.

2 Related Work

A number of models have been proposed in the literature to disambiguate PIEs, with a trend shifting from employing linguistic features to more neural approaches, similar to the rest of the field. Fazly et al. (2009) adopt an unsupervised approach relying on the hypothesis that multiword expressions are more likely to occur in different canonical forms when used literally. Sporleder and Li (2009) propose a generalized method (as opposed to “per-idiom classification”) employing cohesion graphs which initially include all the words in the sentences. They hypothesize that a PIE is used figuratively if the removal of the PIE improves the cohesion. Li and Sporleder (2009) prepares a dataset consisting of high confidences instances found by (Sporleder and Li, 2009) and train a supervised classifier to classify the rest of the instances.

Rajani et al. (2014) use a variety of features including bag of all content words along with their concreteness measures and train a L2 regularized Logistic Regression (L2LR) classifier (Fan et al., 2008). Liu and Hwa (2017) also utilize the cues the context of the PIE provides and adopt an ensemble learning approach based on three different classifiers trained on different representations of the context. Liu and Hwa (2018) propose a “literal usage metric” which quantifies the literalness of PIE. This metric is computed as the average similarity between the words in the sentence and the “literal usage representation” which is the set of the words similar to the literal meanings of the PIE’s main constituent words found in large corpus. Do Dinh et al. (2018) use a multi-task learning approach covering four different non-literal language using tasks including classification of idiomatic use of infinitive-verb compounds in German using recurrent Sluice networks (Ruder et al., 2019). Similar to (Sporleder and Li, 2009), (Liu and Hwa, 2019) adopt a generalized approach and propose a novel “semantic compatibility model” which is a modified version of CBOW, adapted specifically to the disambiguation of the PIEs task.

In a related line of research, contextual embeddings are successfully applied to the general problem of word sense disambiguation (WSD). Wiedemann et al. (2019) show that BERT embeddings form distinct clusters for different senses of a given word in line with its promise to be contextual. Huang et al. (2019) approach WSD as a sentence pair classification task and fine-tune BERT where the input consists of a sentence containing the target word and the one of its glosses and the objective is to classify if the gloss matches the sense of the target word in the sentence.

3 Method

The task here is to distinguish the compositional (literal) and non-compositional (idiomatic) usages of a *known* PIE in a certain context as opposed to MWE extraction which is the task of discovering MWEs in a corpus. Hence, the input to our method is a set of sentences containing a target PIE. We regard disambiguation of PIEs as a word sense disambiguation problem. Our basic assumption is that the context, in which PIEs occur literally and figuratively are distinct enough from each other to be assigned a fundamentally different contextual representations. Below, we briefly introduce the contextual language model we use in the experiments, BERT, followed by the descriptions of the supervised and the unsupervised classifiers.

3.1 BERT

BERT (*Bidirectional Encoder Representations for Transformers*) is a multi-layer Transformer encoder based language model (Devlin et al., 2019). As the transformer encoder reads its input at once, BERT learns words full context (both from left and from right), as opposed to directional models where the input is processed from one direction to another. BERT takes a pair of sentences padded with the special "[CLS]" token in the beginning of the first sentence and "[SEP]" token after the end of each sentence indicating sentence boundaries.

BERT is trained with two objective functions on large-scale unlabeled text: (i) Masked Language Modelling (MLM) and (ii) Next Sentence Prediction (NSP). In MLM, 15% of the input tokens are randomly replaced with a special "[MASK]" token and the task is to predict the masked token by looking at its context. Contrary to the traditional language modelling, where the task is to predict the next word given a sequence of words, the MLM objective forces BERT to consider the context in both sides hence increases its context sensitivity. The NSP objective is a binary classification task to determine if the second sentence in the input follows the first one in the original text. During training, BERT is fed with sentence pairs where half of the time the second sentence is randomly selected from the full corpus.

3.2 Supervised Classification

The supervised model consists of an encoder and a classifier. The task of encoder is to assign each token a representation in a way that every occurrence of each word is represented differently, reflecting their context. We use two different BERT models (Devlin et al., 2019) as encoders in our experiments:

- **Monolingual BERTs** We use bert-base-cased and German-bert[1] as the monolingual BERT models. Each model has the same architecture, consisting of 12 transformers layers and trained on huge monolingual corpus of the respective language.

- **multilingual BERT (mBERT):**[2] mBERT is trained on the concatenation of the 104 Wikipedia dumps with shared word-piece vocabulary. Since the training data does not contain any cross-lingual signal, the source and the extent of the cross-lingual capabilities of mBERT has been a topic of research on its own (Pires et al., 2019).

Since BERT's internal tokenizer splits some words into multiple tokens, e.g. 'microscope' becomes ['micro', '##scope'], we first compute a word-token map which keeps track of the word pieces PIEs are split into. Then, each PIE is represented by the average of their word piece embeddings,

$$V_{PIE_i} = \frac{1}{k} \sum_{j=1}^{k} v_{i,j}$$

where k is the number of word pieces that PIE is split into; $v_{i,j}$ is the representation of the j^{th} word piece in the i^{th} sentence in the dataset. We only count the lexicalised components in the canonical form

[1] https://deepset.ai/german-bert

[2] https://github.com/google-research/bert/blob/master/multilingual.md

of the PIEs as its constituents, e.g. we would leave out the embedding of any realization of *someone* from the embeddings of the MWE *break someone's heart*.

A typical characteristic of compositional PIEs is that their component words display larger variation of inflectional forms than idiomatic PIEs, which is a property that has previously been used as a feature for the purpose of disambiguation (Fazly et al., 2009) (e.g. "broke a leg" can be more likely to be used with the literal sense as opposed to "break a leg" which is almost always used figuratively). Yet, this correlation between the form and the meaning may obscure the results of our experiments as our main aim is to test the degree of contextuality captured by these contextual embeddings. Hence, in order to control for this variation, we lemmatize all the words in PIEs before feeding them to the encoder. In the case of German PIEs, where whether a PIE is written as one word or two words is a strong indicator of its sense, we always spell them as two words. We do not modify the sentence which we pass to encoder in any other way. As for classifier, we use a simple single-layer perceptron to predict the correct usage.s

3.3 Unsupervised Model

The unsupervised model uses the same representations that are used in the supervised setting. We use the hierarchical agglomerative clustering (HAC) algorithm (Day and Edelsbrunner, 1984). We experimented with various configurations and finally adopted Ward as the linkage criterion with Euclidean distances as the similarity metric. Additional experiments with k-means clustering algorithm also yielded similar results but we choose HAC over k-means as it is a deterministic algorithm so the results are more stable.[3]

The unsupervised model relies on the observation that the multiword expressions are semantically in sharp contrast with their surrounding context when used idiomatically, following the previous studies (Peng and Feldman, 2016; Liu and Hwa, 2018). We quantify these heuristics as the average of the cosine similarities between the words in the sentence and the PIE inspired by (Liu and Hwa, 2018):

$$score = \frac{1}{L}\sum_{j=1}^{L}\cos(V_{PIE}, w_j)$$

where w_j is the jth word in the sentence and $\cos(V_{PIE}, w_j)$ is the cosine similarity between the word embedding and the embedding of the PIE. Following our heuristics, we label all PIEs as "idioms" in the cluster, in which the average cosine similarity between PIEs and the sentence they occur in is the lowest.

4 Experiments

We conduct our experiments on the widely used datasets in two languages: the VNC dataset (Cook et al., 2008) and SemEval5b (Korkontzelos et al., 2013) for English and the Horbach dataset for German (Horbach et al., 2016).

Dataset	Language	# of MWEs	Idiom	Literal	Total
VNC	English	12	489 (66.4%)	248 (33.6%)	737
SemEval5b	English	10	1204 (50.7%)	1172 (49.3%)	2376
Horbach	German	6	3369 (64.2%)	1880 (35.8%)	5249

Table 1: Statistics of the datasets used in the experiments. Note that the statistics reflect the subset of the respective dataset used in experiments.

In order to have comparable results, we follow the the official train/test split of Semeval5b dataset whereas for VNC dataset, we used multiword expressions which have at least 10 instances with both literal and idiomatic usage following (Liu and Hwa, 2019). Since there is not any official train/test split for both VNC and Horbach datasets, we report the results of 5-fold cross-validation for the former[4] and 10-fold for the latter. We use Scikit-learn library (Pedregosa et al., 2011) to implement both perceptron

[3] All model selection experiments were performed with the VNC dataset only, thus leaving the larger SemEval5b and Horbach datasets untainted.

[4] Due to the limited size of the VNC dataset.

	Semeval5b		VNC		German Dataset	
Model	Acc	F-score	Acc	F-score	Acc	F-score
(Fazly et al., 2009)†	-	-	0.74	0.73	-	-
(Li and Sporleder, 2009)†	0.62	0.64	0.66	0.67	-	-
(Rajani et al., 2014)	0.75	0.71	0.7	0.69	-	-
(Liu and Hwa, 2017)	0.77	0.77	0.75	0.75	-	-
(Liu and Hwa, 2018)†	0.74	0.75	0.75	0.73	-	-
(Liu and Hwa, 2019)†	0.75	0.76	0.73	0.75	-	-
(Horbach et al., 2016)	-	-	-	-	0.86	-
(Do Dinh et al., 2018)	-	-	-	-	0.88	-
mBERT (Unsupervised)	0.81	0.81	0.69	0.69	0.50	0.55
mBERT (Supervised)	0.91	0.91	0.89	0.85	0.88	0.90
BERT-base (Unsupervised)	0.79	0.78	0.73	0.73	0.55	0.59
BERT-base (Supervised)	**0.94**	**0.93**	**0.91**	**0.90**	**0.94**	**0.94**

Table 2: Averaged results across all idioms in datasets. *BERT-base refers to the monolingual BERT trained on the language of the respective dataset. † indicates an unsupervised baseline.

and agglomerative clustering. The learning rate of the perceptron is set to 1×10^{-5}. The embeddings are normalized before they are fed into the classifiers. As the length of the available context differ for each dataset, we limit the context to the sentence containing the PIE. We use the embeddings from the last layer of the BERT models in the experiments; yet, we conduct a layer-wise analysis as well (see Section 6).

5 Results

Our average results with a detailed comparison with the previous studies are provided in Table 2 and per-idiom results in Figure 1 and in Appendix A. We report the overall accuracy and the F-score for the idiomatic ("figurative") class. The results indicate that contextual embeddings is clearly a better alternative to the previous approaches. The supervised classifier trained on monolingual BERT embeddings achieves the best performances, improving the current state-of-the-art models from 76+% to 91+% F-score on English and from 88% to 92% accuracy on German datasets. Similarly, the unsupervised classification outperforms or is on par with the previous state-of-the-art results on the English datasets but fails to perform equally well on German, which is further discussed in the next section.

Switching to the multilingual contextual embeddings does not lead to a significant decrease in performance, especially in the supervised setting where the results stay considerably above the previous state-of-the-art. It must be noted that the relatively lower performance of the multilingual embeddings in the unsupervised setting is because of a significant drop with certain PIEs, not due to a general failure of the classifier across all PIEs (Figure 1).

6 Discussion

In this section, we further discuss some implications of our results. Overall, we comprehensively evaluated our approach in three datasets. The performance of the supervised classification is pretty consistent across all the PIEs in two languages, ranging between 0.77 to 1.00 F-score with a mean of 0.92 (±0.06). Hence, the increase in the average results are not due to a significant increase in a subset of PIEs but constant improvement in all PIEs covered in the datasets.

As for unsupervised classification, in line with our hypothesis, most of the time BERT embeddings form distinct enough clusters corresponding the different usages of PIE, allowing high performing unsupervised classification. Yet, the unsupervised classifier is more prone to make errors as it completely fails with certain expressions which significantly lowers its overall performance (see Figure 1). We group the errors of the unsupervised classification under two categories:

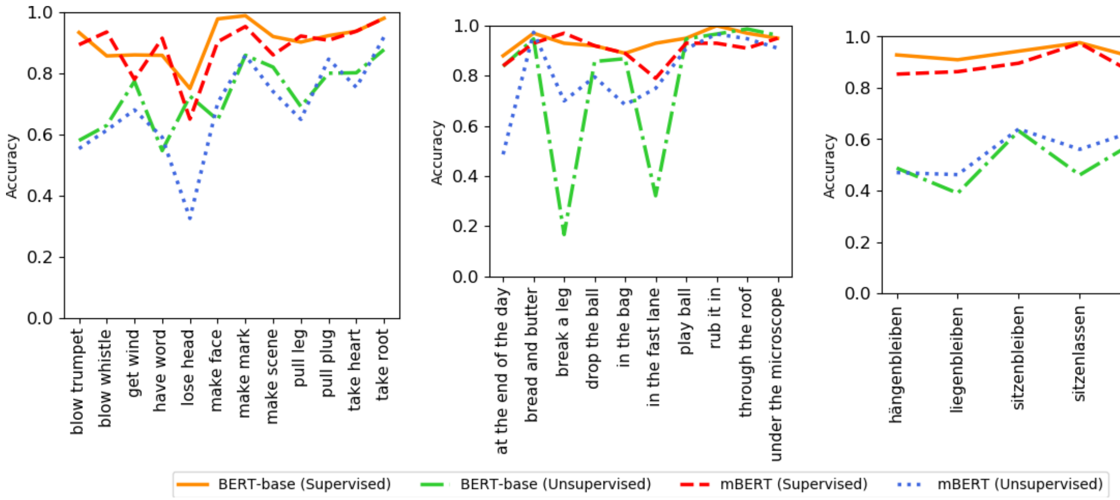
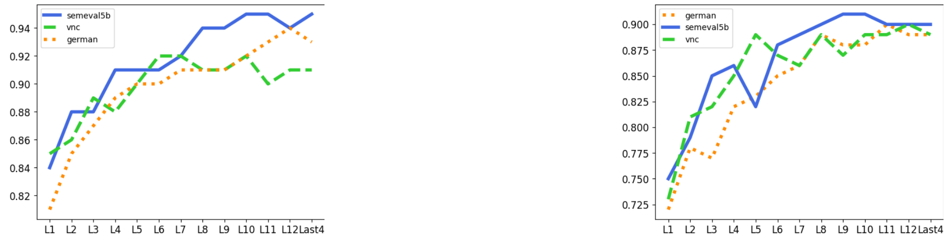


Figure 1: Idiom-wise performance (accuracy) of both classifiers with monolingual and multilingual contextual embeddings. The MWEs are represented in alphabetical order and the lines are added for visibility.

- **Clustering errors** occur due to the formation of poor clusters, consisting of PIEs with different usages. Clustering errors happen relatively rarely in English, where there are only four expressions ("blow whistle", "pull leg", "break a leg", "in the fast lane") with F-score < 0.6; as opposed to German where the unsupervised classifier achieves only 0.59 F-score on average. We suspect that behind the high error rate in German lies the fact that German MWEs exhibit a wider range of polysemy both in literal and figurative interpretation. Horbach et al. (2016) also discusses this point as one of the challenges during annotation, stating that there are not very clearly separated uses of the respective verbs in the dataset, as opposed to, e.g., "bread and butter" in English which has a dominant figurative interpretation. For example, according to (Horbach et al., 2016), stehen+bleiben (stand+still) has a large number of meanings, some of which are *(i) a person's heart may stand still; (ii) people may stand still in their mental development; (iii) you can claim that a statement cannot "remain standing" (remain uncontradicted).* This point is also visible in the dendrograms of German PIEs where there are more distinct clusters on the lower levels (Figure 3). A preliminary analysis of these clusters show tendencies towards this direction, but a more systematic evaluation is left for future work.

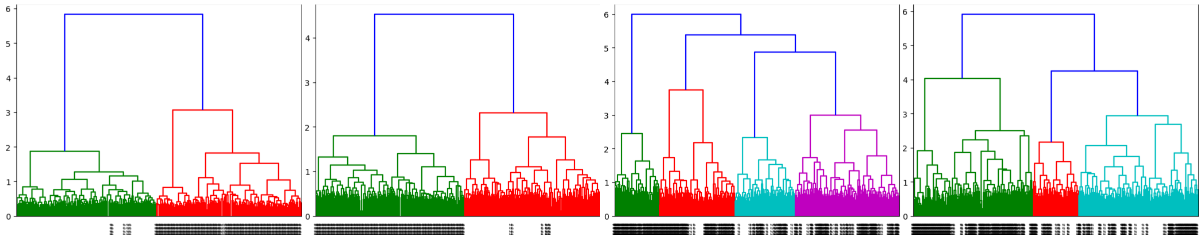
- **Labeling errors** In this case, the lower performance of the unsupervised classifier is due to the failure of our heuristics to label the clusters correctly rather than the formation of poor clusters. The most representative example of this error is the expression "break a leg" where the supervised classification achieves the F-score of 0.89 whereas the unsupervised classifier completely fails as our heuristics fail to label the clusters correctly. We ran a further experiment with an updated heuristics where we directly measure the cosine similarity between the sentence and the PIE by



(a) BERT-base

(b) Multilingual-BERT

Figure 2: The averaged results in accuracy over all layers.



(a) “Bread and butter” (b) “Through the roof” (c) “Liegenbleiben” (d) “Stehenbleiben”

Figure 3: Hierarchical clustering of several cherry-picked English and German PIE embeddings obtained from the respective monolingual BERT model. The leaves corresponding to idiomatic examples are labeled whereas the rest are left empty in order to visualize how the idiomatic and literal instances are separated across clusters.

representing the former as the average of its constituents' embeddings (as opposed to average of the the pair-wise similarity between the PIE and the words in its context). However, the updated heuristics also yielded the same results, highlighting the need for more elaborate heuristics.

Furthermore, we ran all the experiments without performing lemmatization on the target expressions (see Section 3.2) to see if lemmatization had any adverse effect on BERT. Overall, lemmatization turned out to lead to mixed results (1 to 2 point change in F-score) but surprisingly mostly positive; the surface (*unlemmatized*) forms achieve slightly better performance (+1 F-score) only on VNC dataset in the supervised setting and on SemEval dataset in the unsupervised setting when multilingual embeddings are employed. However, as discussed in Section 3.2, without lemmatization it is not possible to know if the classifiers exploit the possible correlation between the surface forms and associated usages. Therefore, we believe lemmatization is a necessary pre-processing step as it allows us for that correlation, without harming the performance.

We, additionally, conducted a layer-wise analysis as different layers of BERT is shown to capture different properties of the language (Tenney et al., 2019). In addition to each layer, we experiment with the concatenation of the last four layers following the original BERT paper (Devlin et al., 2019) which claims that it yields the best contextualized embeddings. The results show that the sixth layer and upwards yield better performances where the concatenation of the last four layers leads mixed results, leading a slight drop on two datasets and increase in one (Figure 2).

Finally, as can be seen in Figure 1, the performance of the supervised classifier with mBERT embeddings are consistent across PIEs which suggests that disambiguation of PIEs can be performed with high accuracy in a large number of languages, requiring only a small set of annotated sentences, e.g. the portion of the VNC dataset used in the experiments contains only 61 sentences annotated per MWE on average (see Table 1).

7 Conclusion

In the current paper, we have proposed two methods, one supervised and one unsupervised, for disambiguation of potentially idiomatic expressions in running text. Our models utilize contextual embeddings which are able to recognize the different usages of the same lexical units and assign representations accordingly. Experimental results in two languages show both of our classifiers substantially outperform the previous state-of-the-art; yet, there is much room for improvement, especially with unsupervised classification which is less stable. The proposed methodology, furthermore, is shown to have a high potential to be extended into a large number of languages thanks to the multilingual contextual embeddings.

Acknowledgements

We would like to thank Mats Wirén for his useful comments and NVIDIA for their GPU grant.

References

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.

William HE Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

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, June. Association for Computational Linguistics.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

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

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

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing 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 3500–3505.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, pages 39–47.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 315–323. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731.

Changsheng Liu and Rebecca Hwa. 2019. A generalized idiom usage recognition model based on semantic compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6738–6745.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Jing Peng and Anna Feldman. 2016. Experiments in idiom recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2752–2761.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Nazneen Fatema Rajani, Edaena Salinas, and Raymond Mooney. 2014. Using abstract context to detect figurative language.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4822–4829.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

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

Appendix A Per-Idiom Results

	Supervised		Unsupervised	
MWE	Acc	F1	Acc	F1
blow trumpet	0.93	0.95	0.58	0.66
blow whistle	0.86	0.82	0.63	0.53
get wind	0.86	0.77	0.77	0.77
have word	0.86	0.92	0.55	0.66
lose head	0.75	0.78	0.72	0.72
make face	0.98	0.98	0.65	0.60
make mark	0.99	0.99	0.86	0.91
make scene	0.92	0.94	0.82	0.82
pull leg	0.90	0.79	0.69	0.55
pull plug	0.92	0.95	0.80	0.81
take heart	0.94	0.96	0.80	0.84
take root	0.98	0.99	0.88	0.92
Average	0.91	0.90	0.73	0.73

(a) VNC dataset

	Supervised		Unsupervised	
MWE	Acc	F1	Acc	F1
at the end of the day	0.88	0.91	0.84	0.86
bread and butter	0.97	0.97	0.95	0.95
break a leg	0.93	0.89	0.17	0.00
drop the ball	0.92	0.86	0.86	0.77
in the bag	0.89	0.90	0.87	0.88
in the fast lane	0.93	0.95	0.32	0.46
play ball	0.95	0.95	0.95	0.95
rub it in	1.00	1.00	0.97	0.98
through the roof	0.97	0.98	0.99	0.99
under the microscope	0.95	0.93	0.96	0.95
Average	0.94	0.93	0.79	0.78

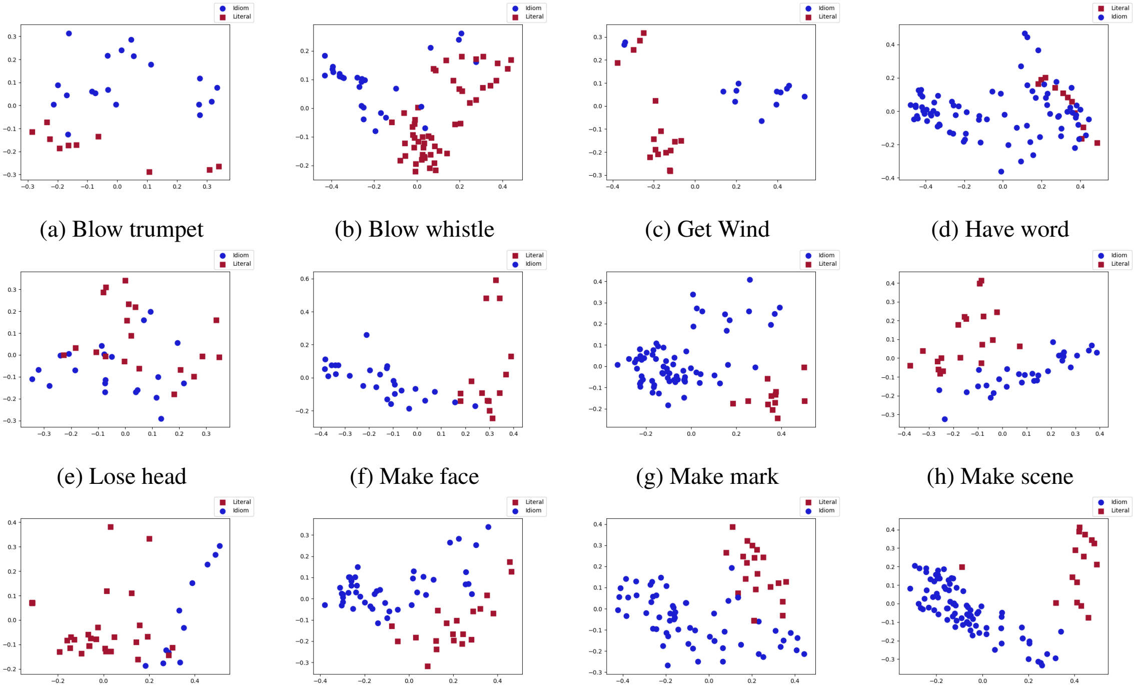
(b) SemEval5b dataset

	Supervised		Unsupervised	
MWE	Acc	F1	Acc	F1
hängenbleiben	0.93	0.95	0.49	0.57
liegenbleiben	0.91	0.91	0.39	0.49
sitzenbleiben	0.94	0.96	0.63	0.67
sitzenlassen	0.98	0.99	0.46	0.60
stehenbleiben	0.91	0.91	0.60	0.59
stehenlassen	0.94	0.93	0.74	0.60
Average	0.94	0.94	0.55	0.59

(c) German dataset

Table 3: Per-idiom results of our supervised and unsupervised classifiers across datasets using monolingual BERT models

Appendix B Visualization of MWE Embeddings



(a) Blow trumpet

(b) Blow whistle

(c) Get Wind

(d) Have word

(e) Lose head

(f) Make face

(g) Make mark

(h) Make scene

(i) Pull leg

(j) Pull plug

(k) Take heart

(l) Take root

Figure 4: PCA plots of the BERT-base embeddings for the VNC dataset