How does BERT capture semantics? A closer look at polysemous words

David Yenicelik
ETH Zürich
yedavid@ethz.ch

Florian Schmidt
ETH Zürich
florian.schmidt@inf.ethz.ch

Yannic Kilcher
ETH Zürich
yannic.kilcher@inf.ethz.ch

Abstract

The recent paradigm shift to contextual word embeddings has seen tremendous success across a wide range of down-stream tasks. However, little is known on how the emergent relation of context and semantics manifests geometrically. We investigate polysemous words as one particularly prominent instance of semantic organization. Our rigorous quantitative analysis of linear separability and cluster organization in embedding vectors produced by BERT shows that semantics do not surface as isolated clusters but form seamless structures, tightly coupled with sentiment and syntax.

1 Introduction

Word embeddings have not only proven to be excellent representations in standalone tasks (Mikolov et al., 2013; Pennington et al., 2014; Wang et al., 2019) but have revolutionized the way modern NLP architectures are built (Collobert et al., 2011), and by now encode text input for virtually every task available. Recently, this approach has been paired with the transformer architecture (Vaswani et al., 2017) and a selection of pre-training tasks to bootstrap more powerful *contextual* word embeddings such as the ones produced by BERT (Devlin et al., 2018). s

The paradigm of encoding a word in its context has elevated the embedding methodology once more and from several perspectives. First, performance improvements on down-stream tasks are extraordinary across a wide range of tasks (Ethayarajh, 2019; Devlin et al., 2018; Wang et al., 2018). Second, the embedding space now must incorporate a vastly larger number of vectors, and its organization becomes an interesting research question on its own, especially given the largely unattributed performance gains.

In this work, we investigate the important concept of *polysemy* as one prominent example of semantic sub-space organization. Given that a word such as 'bank' can have several meanings, how are the corresponding vectors arranged in a contextual embedding space?

We investigate the organization of polysemous words in BERT embeddings through the concepts of separability and clusterability using the WordNet annotations in SemCor (Miller et al., 1990). Our particular focus is a rigorous quantitative rather than purely qualitative analysis.

2 Related Work

Work connecting polysemy and word vector representations is often limited to static word embeddings where context has to be re-introduced through graph-based approaches (Remus and Biemann, 2018), auxiliary corpora (Pelevina et al., 2016), or even image data (Bruni et al., 2013). Usually, word sense disambiguation (WSD) performance is then chosen as a proxy to semantic disambiguation (Pilehvar and Camacho-Collados, 2019), yet no insights into the organization of the vector space are obtained. In a similar spirit, Kageback and Salomonsson (2016) add context through a recurrent encoder, yet do not analyze the geometry of these encodings.

In the meantime, the WSD task has been tackled successfully with BERT embeddings and Wiedemann et al. (2019) show that even a non-parametric approach suffices, which confirms that BERT must arrange word vectors according to semantic properties and suggests that no additional semantic pre-training is necessary (Levine et al., 2019).

When BERT embeddings of a polysemous word are analyzed, the findings are often summarized as a Silhouette score (Rousseeuw, 1987) or custom

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 156–162
Online, November 20, 2020.

variance measures (Ethayarajh, 2019). While this allows to compare the average displacement due to semantic change across words, it does not give us a good sense of the overall structure of word vectors. In addition, the embedding space produced by BERT has been analyzed in terms of syntactic features, such as parse-trees (Coenen et al., 2019; Jawahar et al., 2019), part-of-speech, verbs and arguments (Shi et al., 2019; Ribeiro et al., 2019).

It is clear that BERT distinguishes polysemous words at least locally by nearest neighbors (Schmidt and Hofmann, 2020). However, the extent to which clusters are formed and how they are connected has only been addressed qualitatively (Coenen et al., 2019; Jawahar et al., 2019; Wiedemann et al., 2019), and no agreed-upon answer has emerged. This can be partly attributed to their qualitative methodology.

3 Method

How can we verify a hypothesis about the organization of polysemous words without manually inspecting the geometry for each word? Given a set of sentences with annotated polysemy, two strategies emerge: First, we can inspect the embedding space through the lens of a classifier with a clearly defined hypothesis set and take its accuracy as a signifier for the corresponding organization. Second, we can use an unsupervised approach to detect sub-space organization and compare its result to the WordNet labels using an appropriate similarity metric. We will proceed by analyzing both questions.
In our experiments, we consider the output of the last layer of BERT as the contextual word embeddings since this layer is most commonly used for downstream tasks, as depicted in Figure 1. To work with a discrete formalization of semantics, we use the WordNet 3.0 annotations in the SemCor 3.0 sentence dataset (Miller et al., 1990). This allows us to retrieve embeddings that are annotated with a ground-truth *semantic class* label. SemCor is one of the largest sense-annotated corpora with 37,176 sentences, enabling us to quantify semantics and sample labelled word embeddings.

3.1 Linear Separability

Before turning to the clustering task, we investigate to what degree semantic classes can be separated by a hyperplane in embedding space, resulting in *semantic regions* as depicted in Figure 2. To this end, we train a simple linear classifier on top of the BERT embeddings (without fine-tuning) to predict the semantic class and report accuracy. Crucially, we down-project the 768-dimensional vectors using PCA ensuring that separability is not merely a consequence of high dimensionality. In the high-dimensional setting, this allows to assess to what extent semantic regions do form.

3.2 Clusterability

Once the presence of semantic regions has been concluded, we want to investigate the extent to which clusters form and how they are connected. For this, we train clustering models to understand the modality of the data, and to what extent clusters are in isolation from each other.

Because we are interested in practical gains, we refrain from using purely theoretical tools and clusterability scores (Ackerman and Ben-David, 2009; Mccarthy et al., 2016). In contrast, we use interpretable clustering models (Frey and Dueck, 2007; Ester et al., 1996; Campello et al., 2013; Comani-

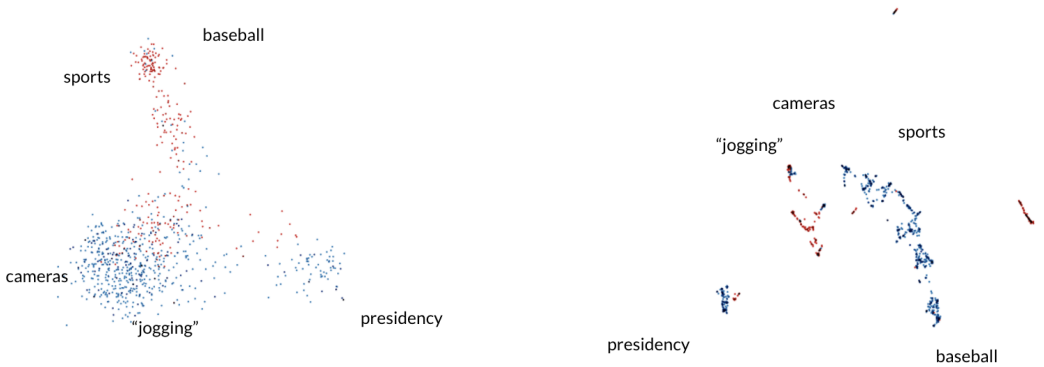

Figure 1: PCA (left) and UMAP (right) visualizations for contextual word embeddings sampled for the word `run`. Red points denote *nouns*, blue points denote *verbs*

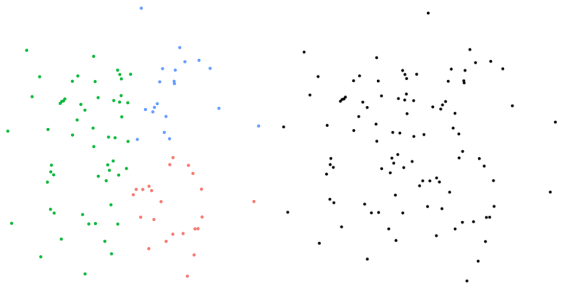

Figure 2: Datapoints generated by sampling from a normal distribution. (left) includes class information denoted by the green, blue and red colors, forming semantic regions. This data is linearly separable, as there is always a hyperplane separating any two classes. (right) The class label information is not available. It is unclear how a clustering would look like.

ciu and Meer, 2002) that can detect the number of clusters in the data, as well as an adapted version of the Chinese Whispers algorithm (Biemann, 2006) that accounts for the *hubness property* [1] amongst embedding vectors outputted by BERT. The Chinese Whispers algorithm relies on a graph produced by the word embeddings and identifies clusters by passing messages between the nodes of the graph. From the sampled word embeddings we create the graph adjacency matrix M by calculating the pairwise cosine similarity between embeddings, and similar to Ribeiro et al. (2019), prune any edges which correspond to a cosine similarity lower than $w_{\text{cutoff}} = \mu(M) + c\sigma(M)$ where c is a hyperparameter, and μ and σ are the mean and standard deviation of all cosine similarities recorded in M. Hubs are defined as the top n embedding vectors with highest cumulative cosine similarities.
The development and test sets consist of $\frac{n}{2}$ words respectively, including their set of sampled embedding vectors.

To score the overlap between a predicted clustering and the underlying ground-truth labels, we use the Adjusted Random Index (ARI) (Rand, 1971; Hubert and Arabie, 1985), which returns a similarity measure where a value of 1 implies an identical clustering up to a permutation and a value of 0 implies random predictions. Please note that this also introduces a small penalty when more clusters are introduced than actually present in the dataset according to the cluster-class-labels. However, preventing this is not in the scope of this work, and as such we do not further investigate this.

If no clustering is found, one can say with high confidence that the semantic regions are not occurring in different modes, and rather transition seamlessly into one another. This allows for an assessment in high dimensional space to what extent semantic regions are obvious, apparent by distinct modes. The motivation behind both experiments is visually depicted in Figure 2.

4 Experiments

We proceed with a discussion of the experimental results.

4.1 Bias in SemCor

First we aim to develop an understanding of bias in SemCor, as any bias in the data will propagate on to further observations. We conduct a simple experiment where we analyse the distribution of occurrences of semantic class ids.

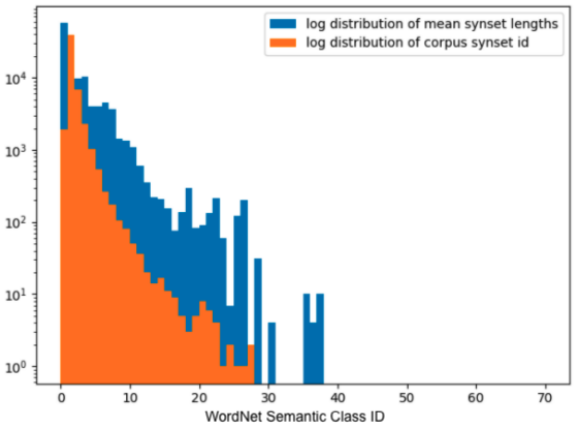

Figure 3: A cumulative plot over all words with WordNet senses within SemCor 3.0 and their respective frequencies. The SemCor data is biased. Words with a low WordNet sense index, i.e. close to 0, occur more often than words with a high WordNet sense index, i.e. above 5. There would be no bias if the two distributions would overlap. The skew could be a natural effect of how words with lower WordNet indices are assigned to more frequently used words.

Figure 3 depicts that the SemCor corpus is biased towards semantic classes which have a lower WordNet class ID. This could be due to the nature of WordNet, likely assigning low id indices to frequently used words. This requires us to oversample underrepresented classes for select experiments.

[1] *hubs* are embeddings close to a majority of other embedding vectors, degrading performance (Conneau et al., 2017)

4.2 Linear Separability

We now turn our attention to what extent closed semantic regions exist in the embedding space. For a fixed word w that frequently occurs in SemCor, we sample up to $n = 500$ embedding vectors, apply 5-fold cross-validation, and oversample any imbalanced-class datasamples. The input is normalized, and we apply dimensionality reduction using PCA to k components, ensuring that each of the semantic classes contains at least 20 samples in the dataset. We use the SemCor semantic class labels as the response variables for the classification task. We only include semantic classes for the given word, leaving us with few class-labels. Results are shown in Table 1.

k	% variance	accuracy (mean / std)
10	0.30	0.74 / 0.05
20	0.44	0.80 / 0.04
30	0.54	0.82 / 0.03
50	0.70	0.87 / 0.04
75	0.79	0.83 / 0.04
100	0.85	0.89 / 0.03

Table 1: Average mean and standard deviation of the accuracy of a linear classifier trained on the 2 most common semantic classes for the words `was`, `one`, `is`. The choice of words is limited to the datasize of SemCor to allow for a significant size of datasamples.

Accuracy rates of over 75% are achieved with $k = 20$. The % *variance* refers to the explainable variance when the largest k eigenvalues are kept, as calculated by $\sum_i^k \sigma_i$ where σ_i is the ith largest eigenvalue, hinting to how much information according to the largest k principal components are kept. Similar results are achieved for 2-class and multi-class classification tasks with other words (see Appendix A.2). We conclude that the individual semantic classes are – to a reasonable extent – linearly separable. As such, contextual word embeddings are not randomly distributed over the embedding space, and closed semantic regions do form.

4.3 Polysemy vs. Variance

Before we analyze the structure of individual semantic classes, we want to understand how polysemy relates to the mean standard deviation of all contextual word embeddings X sampled for the word w. This helps us to understand how we need to adapt different clustering models.

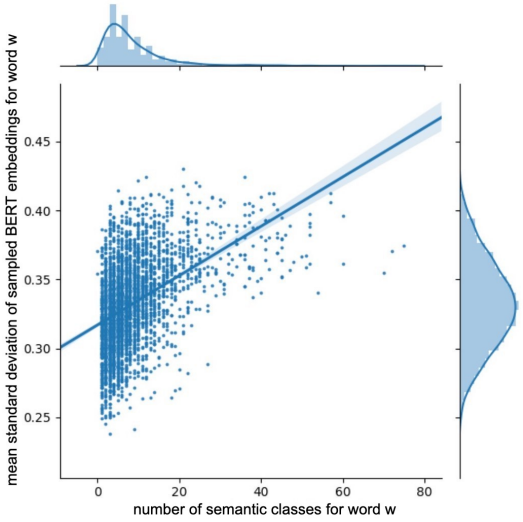

Figure 4: For each word w, we sample up to $n = 500$ contextual word embeddings X. We calculate the mean standard-deviation across embedding-dimensions as $\sum_i^n \sum_j^d {x_i}^j$ where $x_i^j \in \mathbf{R}^d$ is the jth dimension of the ith sampled embedding vector for word w. WordNet is used to retrieve the number of semantic classes of w, denoting the amount of polysemy of word w.

Figure 4 shows that polysemous words have high variance, an idea initially put forth by Miller and Charles (1991). As such, vectors of polysemous words seem to be distributed at least as dispersed around the space as non-polysemous words do. Notice that the converse is not true, as there are non-polysemous words that have high variance. Amongst others, these could include stopwords as hinted by Ethayarajh (2019).

4.4 Clusterability

We now want to understand to what extent distinct semantic clusters exist. For a set of words $w_1, \ldots, w_n$, we sample up to $n = 500$ embedding vectors per word from SemCor and the `news.2007.corpus` [2] and apply dimensionality reduction using PCA to k dimensions. Due to the limited size of SemCor, we set the words in the development set to `was`, `thought`, `made`, `only`, `central`, `pizza` and the set of words in the test set to `run`, `round`, `down`, `bank`, `key`, `arms`. We include both polysemous words, as well as words which have a single recorded WordNet meaning, such that our experiments do not overfit to polysemous words. With default-package hyperparameters, all clustering algorithms would indicate that no distinct clustering could be found,

[2] http://www.statmt.org/wmt14/training-monolingual-news-crawl/

i.e. the sampled word embeddings would form a continuous density. Because we want to see to what extent BERT conforms to commonly accepted linguistic senses as given by WordNet, we apply the NetworkX Chinese Whispers implementation (Hagberg et al., 2006) on the resulting graph. Hyperparameters and their respective bounds for all clustering models are listed in Appendix A.1. We include the `[SEP]` tag at the end of the sentence, as this increases performance on all clustering methods. The ARI is exclusively calculated on samples for which we have the ground-truth cluster label and stems from the mean of multiple such word clusterings. We notice that choosing suitable hyperparameters is non-trivial and thus apply automated model- and hyperparameter selection, making use of random search (Bergstra and Bengio, 2012) and bayesian optimization (Wang et al., 2013) [3].

Clustering Model	ARI Score
Affinity Propagation	0.316
Modified Chinese Whispers	**0.457**
DBScan	0.170
HDBScan	0.298
MeanShift	0.251

Table 2: The maximum ARI scores achieved during hyperparameter optimization on different models for $k = 20$ and $n = 1000$.

The models used and their maximal performance after 300 trials of hyperparameters search are recorded in Table 2. Our modified Chinese Whispers algorithm is the best-performing clustering model. However, with an ARI score of 0.457, this method is not able to perfectly distinguish between multiple WordNet semantic classes [4]. To understand why this is the case, we proceed with a qualitative evaluation of some resulting clusters. One such clustering is depicted in Table 3, presenting four partitions for `arms` [5]. We achieve similar such results for 9 other words but focus on one example for conciseness. Notice that the clusters differ not only in semantics but also in other linguistic phenomena, most notably sentiment.

Given the quantitative and qualitative evaluation, we conclude that one cannot generalize that a clear distinction between semantic concepts in contextual word embeddings produced by BERT exists. One of numerous counterexamples is underlined in the left visualization of Figure 1. Certain combinations of semantics, syntax, and sentiment are more frequent than others (Hagoort, 2003; May et al., 2019), likely affecting the subspace structure and sometimes resulting in clusters that are distinct due to their simultaneous difference in both semantic and syntactic features (see Appendix A.3). However, this work also poses the question to what extent rule-based and handcrafted notions of semantics, such as the ones given by WordNet, are appropriate, opening the question to what extent BERT actually encodes a more flexible notion of semantics that is not rooted in hard distinctions between senses. We leave analysis in this direction to future work.

Partition	Representative Sample
1	Ms. Gotbaum tried to slide her handcuffed **arms** from her back to her front ...
2	She swooped him up into her **arms** and kissed him madly ...
3	... and shuttle robotic **arms** of a solar array and truss ...
4	The classic years of the **arms** race, the 1950s and '60s before ...

Table 3: Representative samples for the clusters found by the best performing clustering model for the word `arms`. Partitions 1-3 consider a person's arms, whereas partition 4 considers `arms` as a synonym to `weaponry`. Partitions 1, 2 and 3 strongly contrast in sentiment (scared, loving, and confident respectively).

5 Conclusion

In this paper, we investigated how contextual word embeddings produced by BERT capture semantic concepts with a strong focus on polysemy. Our findings show that BERT creates closed semantic regions that are not clearly distinguishable from each other, seamlessly transitioning from one into another. We have shown that subspace organization is not purely determined by semantics. Instead, it is also intertwined with concepts such as syntax and sentiment. Finally, the repeated limitations of hard distinctions between senses as given via WordNet also open up the question to what extent BERT adds a more flexible notion of semantics, compared to the hard-coded examples formed by

[3] We use the implementation by https://github.com/facebook/Ax

[4] An ARI score of at least 0.7 is desirable to conclude a significant overlap between two clusters

[5] See Table 8 for a complete example clustering

linguists. A better understanding of these relations will be key to developing more interpretable and expressive word embeddings, as well as linguistic knowledge representations.

Acknowledgments

David Yenicelik would like to thank Jason Lee (NYU) for the support setting an initial research question investigating on the structure of contextual word embeddings used in translation tasks, Jeremy Scheurer and Gertrude Yenicelik for discussions and comments, as well as Prof. Thomas Hofmann (ETH Zürich) for valuable discussions, guidance and enabling to conduct this project as part of the Master's thesis. Finally, the authors thank the anonymous reviewers for their constructive feedback.

References

Margareta Ackerman and Shai Ben-David. 2009. Clusterability: A theoretical study. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR 5:1-8.*

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*, pages 281–305.

Chris Biemann. 2006. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. *Workshop on Graph-based Methods for Natural Language Processing*, pages 73–80.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2013. Multimodal distributional semantics. *Journal of Artificial Intelligence Research 49*, pages 1–47.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clusteering based on hierarchical density estimates. *PAKDD 2013: Advances in Knowledge Discovery and Data Mining*, pages 160–172.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019).*

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

Dorin Comaniciu and Peter Meer. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 603–619.

Alexis Conneau, Guillaume Lample, Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *ICLR 2018.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matrin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters. *KDD-96 Proceedings.*

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *EMNLP 2019.*

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science 315 (5814)*, pages 972–976.

Aric Hagberg, Dan Schult, and Pieter Swart. 2006. Networkx. https://github.com/networkx/networkx. Accessed: 2020-05-07.

Peter Hagoort. 2003. Interplay between syntax and semantics during sentence comprehension: Erp effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience 15:6*, pages 883–899.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, pages 193–218.

Ganesh Jawahar, Benoit Sagot, and Djame Seddah. 2019. What does BERT learn about the structure of language? *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics*, pages 3651–3657.

Mikael Kageback and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*, pages 51–56.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. SenseBERT: Driving some sense into BERT. *ACL 2020*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv:1903.10561*.

Diana Mccarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics, Volume 42, Issue 2 - June 2016*, pages 245–275.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, pages 235–244.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes 6*, pages 1–28.

Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. *Proceedings ofthe 1st Workshop on Representation Learning for NLP*, pages 174–183.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. *EMNLP*, page 1532–1543.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. *NAACL 2019*.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, pages 846–850.

Steffen Remus and Chris Biemann. 2018. Retrofitting word representations for unsupervised sense aware word similarities. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Eugénio Ribeiro, Vânia Mendonça, Ricardo Ribeiro, David Martins e Matos, Alberto Sardinha, Ana Lucia Santos, and Luísa Coheur. 2019. L 2 f/inesc-id at semeval-2019 task 2: Unsupervised lexical semantic frame induction using contextualized word representations. *Proceedings ofthe 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 130–136.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics 20*, pages 53–65.

Florian Schmidt and Thomas Hofmann. 2020. Bert as a teacher: Contextual embeddings for sequence-level reward. *ArXiv*, abs/2003.02738.

Peng Shi, Jimmy Lin, and David R Cheriton. 2019. Simple BERT models for relation extraction and semantic role labeling. *arXiv:1904.05255*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *CoRR*, abs/1901.09785.

Ziyu Wang, Masrour Zoghi†, Frank Hutter, David Matheson, and Nando de Freitas. 2013. Bayesian optimization in high dimensions via random embeddings. *AAAI Publications, Twenty-Third International Joint Conference on Artificial Intelligence*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *Conference on Natural Language Processing (KONVENS) 2019*.