On the Relation Between Fine-Tuning, Topological Properties, and Task Performance in Sense-Enhanced Embeddings

Deniz Ekin Yavas[1], **Timothée Bernard**[2], **Benoît Crabbé**[2], **Laura Kallmeyer**[1]
Heinrich Heine University Düsseldorf[1], Université Paris Cité[2]
{deniz.yavas, laura.kallmeyer}@hhu.de[1]
{timothee.bernard, benoit.crabbe}@u-paris.fr[2]

Abstract

Topological properties of embeddings, such as isotropy and uniformity, are closely linked to their expressiveness, and improving these properties enhances the embeddings' ability to capture nuanced semantic distinctions. However, fine-tuning can reduce the expressiveness of the embeddings of language models. This study investigates the relation between fine-tuning, topology of the embedding space, and task performance in the context of sense knowledge enhancement, focusing on identifying the topological properties that contribute to the success of sense-enhanced embeddings. We experiment with two fine-tuning methods: *Supervised Contrastive Learning* (*SCL*) and *Supervised Predictive Learning* (*SPL*). Our results show that SPL, the most standard approach, exhibits varying effectiveness depending on the language model and is inconsistent in producing successful sense-enhanced embeddings. In contrast, SCL achieves this consistently. Furthermore, while the embeddings with only increased *sense-alignment* show reduced task performance, those that also exhibit high *isotropy* and balance *uniformity* with *sense-alignment* achieve the best results. Additionally, our findings indicate that supervised and unsupervised tasks benefit from these topological properties to varying degrees.

1 Introduction

The contextualized word embeddings of pre-trained language models (LMs) are powerful tools, but do not align well with word senses: the embedding spaces typically do not exhibit distinct clusters corresponding to similar meaning (Yenicelik et al., 2020). Therefore, using these embeddings as such does not always yield satisfactory results on word sense identification tasks (Pilehvar and Camacho-Collados, 2019; Samih and Kallmeyer, 2023; Yavas, 2024; Yavas et al., 2024). One way to overcome this problem is to enhance the embeddings with external sense knowledge (Peters et al., 2019; Lauscher et al., 2020; Garí Soler and Apidianaki, 2020; Levine et al., 2020; Mosolova et al., 2024).

Fine-tuning is a widely adopted method for *knowledge enhancement*, where LM embeddings are updated by incorporating external knowledge.[1] However, fine-tuning can also reduce the expressiveness of the LM embeddings, i.e., their ability to capture and distinguish nuanced relations, due to reduced isotropy (Zhang et al., 2022). Therefore, it is essential to determine the effects of fine-tuning on the topology of the embedding space in the context of sense knowledge enhancement. Furthermore, it is also important to understand the relation between topological properties and task performance. Despite their importance, these relations remain unexplored, and this study aims to fill this gap. To this end, we compare different fine-tuning methods, investigate how they transform the embedding space of LMs, evaluate the resulting embeddings in terms of task performance, and finally, identify the topological properties associated with successful sense-enhanced embeddings.

We evaluate the embeddings based on three topological properties: *sense-alignment*, *uniformity* (Wang and Isola, 2020), and *isotropy* (Mu and Viswanath, 2018). *Sense-alignment* evaluates how close to each other are the embeddings of word occurrences sharing the same sense. *Isotropy* evaluates how evenly the embeddings are distributed across all directions in the embedding space, while *uniformity* measures the density of their distribution, evaluating whether they spread uniformly in the embedding space.

Both isotropy and uniformity are related to the expressiveness of the embeddings (Gao et al., 2021; Rajaee and Pilehvar, 2021a; Zhang et al., 2022). Ideally, an embedding space should be uniform and isotropic, while maintaining semantic relations

[1] See the survey by Hu et al. (2023).

as much as possible. This enables utilization of the entire space and directions to capture nuanced semantic relations and differences between word occurrences. However, LM embeddings have inherently limited expressiveness (Gao et al., 2019). As a result, even the embeddings of random words often exhibit high similarity (Ethayarajh, 2019). However, improving the uniformity and isotropy of the LM embeddings can lead to better performance on semantics-related tasks (Biś et al., 2021; Gao et al., 2021; Liu et al., 2021a,b; Rajaee and Pilehvar, 2021a).

We experiment with two fine-tuning methods: *Supervised Contrastive Learning (SCL)* and *Supervised Predictive Learning (SPL)*. They differ primarily in their objectives: SCL aims to push instances of the same class (i.e., sense) together while pushing instances from different classes apart; SPL aims to predict the correct class of an instance. We are particularly interested in these methods because SPL is the standard fine-tuning method for LMs, while fine-tuning via Contrastive Learning has been shown to improve the isotropy and uniformity of the embedding space (see Gao et al., 2021; Liu et al., 2021a,b; Yan et al., 2021, which, however, do not focus on knowledge enhancement).[2]

The few studies that compare SPL and SCL have primarily focused on performance differences, showing better performance with SCL in different contexts, but without investigating the quality of the resulting embeddings (Gunel et al., 2020; Khosla et al., 2020). In the context of sense knowledge enhancement, we aim to investigate whether the performance advantage of SCL holds and to understand the underlying factors in terms of topological properties in performance differences. For this purpose, we conduct detailed experiments, including experiments focusing on the temperature parameter of the loss functions used, as this parameter influences class separation and has been shown to impact the topological properties of the embeddings in Unsupervised Contrastive Learning (UCL; Wang and Liu, 2021).

We enhance the embeddings of LMs with WordNet supersense information (Fellbaum, 1998) using different fine-tuning methods, and evaluate the word sense identification performance with the resulting embeddings on the Word-in-Context task (WiC; Pilehvar and Camacho-Collados, 2019). WiC is a binary classification task, where the goal is to determine if a target word form (e.g., *bass*) has the same meaning in two different contexts (e.g., *she plays the bass* and *this fish is a bass*). We experiment with both supervised and unsupervised setups for the task. This allows us to investigate whether supervised and unsupervised tasks benefit from different topological properties in embeddings to achieve good performance.

In addition to these fine-tuning methods, we conduct further experiments to investigate the effects of two factors on task performance: *task similarity*, and *isotropy in isolation*. For task similarity, we fine-tune the LM on a task similar to the downstream task (WiC). We refer to this fine-tuning approach as *task adaptation*. For isotropy in isolation, we apply the isotropization post-processing method proposed by Mu and Viswanath (2018) to the original LM embeddings, allowing us to investigate the effects of isotropy without incorporating external knowledge to the embeddings. We further evaluate the embeddings obtained from both methods.

Our study provides novel insights into the relation between fine-tuning, topology of the embedding space, and task performance in the context of sense knowledge enhancement. The key findings of this work are as follows.

- *SPL* and *SCL* fine-tuning methods transform the embedding spaces of the LMs differently with respect to sense-alignment, uniformity, and isotropy.
- The properties of the embeddings obtained after fine-tuning appear consistently, regardless of the original embeddings.
- The effectiveness of *SPL* shows mixed results, varying depending on the LM. The embeddings created via SPL do not outperform the original LM embeddings in some cases. In contrast, *SCL* shows more consistent improvements and often achieves the highest scores, making it a more effective method for sense knowledge enhancement.
- The embeddings with only increased *sense-alignment* perform worse than the original LM embeddings. The highest performance is achieved with those that exhibit a significant increase in both *isotropy* and *sense-alignment*.
- The embeddings that achieve the best performances overall balance uniformity and sense-

[2]These studies mostly focus on Unsupervised Contrastive Learning, a related but distinct concept that does not rely on explicit external knowledge.

alignment.

- The relative importance of these properties varies across supervised and unsupervised setups.

2 Related Work

2.1 Enhancing LM Embeddings with Sense Knowledge

Prior research on sense knowledge enhancement has primarily focused on injecting external knowledge during the pre-training phase of the LMs (Peters et al., 2019; Lauscher et al., 2020; Levine et al., 2020). A few studies have fine-tuned LMs for a similar purpose (Garí Soler and Apidianaki, 2020; Mosolova et al., 2024). Garí Soler and Apidianaki (2020) fine-tune BERT on semantic similarity datasets using both a classification head (via SPL) and cosine distance head, which utilizes Cosine Embedding Loss to adjust the embedding distances according to their meaning similarity. They show improved performance on the Graded Word Similarity in Context task (Armendariz et al., 2020) over the original LM embeddings. Mosolova et al. (2024) fine-tune BERT via SCL using examples of use of different senses of the same word form. They show improved performance on the WiC task over the original LM embeddings. Despite the popularity of SPL in other contexts, these studies use other methods than SPL alone. It is still unclear whether SPL can produce similar results, how the two methods differently alter the embedding space, and which topological properties may be associated with any performance differences.

Finally, Bihani and Rayz (2021) introduce a post-processing method to generate isotropic sense-enhanced embeddings. Their approach integrates and adapts the isotropization method of Mu and Viswanath (2018) with the retrofitting method proposed by Faruqui et al. (2015). However, they do not evaluate the resulting embeddings in terms of task performance.

2.2 Topology of the Embedding Space

2.2.1 Alignment and Uniformity

Wang and Isola (2020) propose metrics *alignment*[3] and *uniformity*. They compare SPL and UCL for pre-training encoders for language and vision and show that the embeddings produced by these methods differ in terms of alignment and uniformity, as well as downstream task performance. They find that UCL produces more uniform embeddings, leading to improved performance. Furthermore, they emphasize the importance of balancing alignment and uniformity to generate successful embeddings in terms of downstream task performance.

A similar effect of uniformity has been observed with BERT embeddings specifically. Gao et al. (2021) focus on improving the sentence embeddings of BERT and show that fine-tuning the model via UCL or SCL results in embeddings exhibiting increased uniformity and leading to better results on Semantic Textual Similarity Tasks.

Finally, Wang and Liu (2021) have shown a connection between the temperature parameter in the loss and both the *uniformity* and *tolerance* of the embeddings in UCL. Tolerance is measured as the average similarity of instances within the same class. In their experiments, successful embeddings are only obtained when balancing these two properties.

2.2.2 Isotropy

Previous studies have observed that the embeddings of LMs are anisotropic, and it has been suggested that isotropy is linked to better performance on semantics-related tasks (Biś et al., 2021; Gao et al., 2021; Liu et al., 2021a,b; Rajaee and Pilehvar, 2021a; Yan et al., 2021). Several methods have been proposed to improve isotropy, including regularization techniques during the pre-training of the LM (Gao et al., 2019; Wang et al., 2019a,b; Zhang et al., 2020) and post-processing methods (Bihani and Rayz, 2021; Biś et al., 2021; Rajaee and Pilehvar, 2021a). In addition, fine-tuning LMs via Contrastive Learning has been shown to increase the isotropy of their embeddings (Gao et al., 2021; Liu et al., 2021a,b; Yan et al., 2021).

Anisotropy is also seen in fine-tuned LM embeddings (Rajaee and Pilehvar, 2021b; Zhang et al., 2022) and fine-tuning has been shown to increase anisotropy (Zhang et al., 2022). Furthermore, post-processing methods designed to increase the isotropy of pre-trained LM embeddings tend to be detrimental in terms of downstream task performance when applied to the embeddings produced by fine-tuned LMs (Rajaee and Pilehvar, 2021b; Zhang et al., 2022).

[3]They propose alignment in the context of UCL. The aim of UCL is usually to push the embeddings of a data instance and of its augmented (as in "data augmentation") version closer, while pushing the embeddings of other instances apart. Alignment quantifies the closeness between a data point and its augmented version.

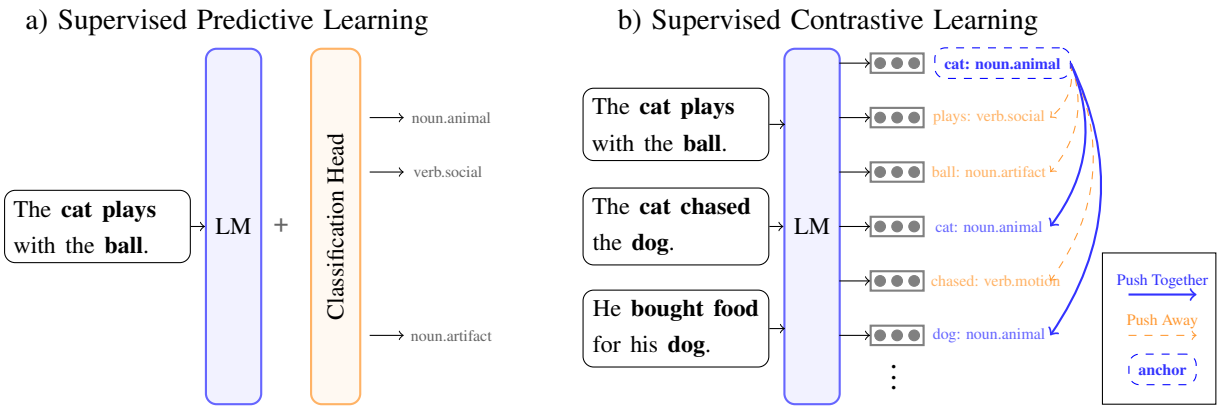

Figure 1: Two fine-tuning methods for sense knowledge enhancement based of WordNet supersenses. For Supervised Predictive Learning (a), a supersense classification head is added to the encoder. For Supervised Contrastive Learning (b), for each *anchor* (all nouns and verbs occurrences in a batch of sentences are simultaneously anchors), the encoder is fine-tuned to push closer to the anchor the word occurrences with the same supersense while pushing away from it those with a different supersense.

3 Methodology

In this section, we describe our external source for sense knowledge (Section 3.1) and the three methods that we use to enhance the contextualized word embeddings produced by two LMs: BERT (*base-uncased*) (Devlin et al., 2019) and RoBERTa (*base*) (Liu et al., 2019).

These methods are (1) fine-tuning the LMs for sense knowledge enhancement using SPL and SCL (Section 3.2), (2) fine-tuning the LMs for task adaptation (Section 3.3), and (3) applying an isotropization post-processing method to the original LM embeddings (Section 3.4). The resulting sets of embeddings are evaluated in Section 4. To prevent over-fitting during fine-tuning (first two methods), the first 8 layers of the LMs are frozen[4] and a dropout rate of 0.7 is used after each hidden layer.[5]

3.1 Data

To fine-tune the LMs for both sense knowledge and task adaptation, we use the SemCor corpus (Miller et al., 1993), which is based on WordNet (Fellbaum, 1998). SemCor provides sentences in which word occurrences that belong to certain parts-of-speech are annotated with a WordNet synset label, which in turn is linked to a *supersense*. Supersenses are broad semantic categories that group word senses and are easier to identify than word-specific senses. For example, "bass" maps to the supersense `artifact` in "she plays the *bass*" and to `animal` in "this fish is a *bass*". Since the WiC task focuses only on verbs and nouns, we ignore all other parts-of-speech. We split the corpus into train, validation, and test sets (70:15:15) and use the test set to evaluate the topology of the embedding spaces, as detailed in Section 4.1.

3.2 Fine-Tuning for Sense Knowledge Enhancement

For Supervised Predictive Learning (SPL), a classification head is added to the LM and the model is fine-tuned by minimization of the Cross-Entropy loss to predict the supersense label of each verb and noun token.[6] See Figure 1, (a), for an illustration.

For Supervised Contrastive Learning (SCL), the LM is fine-tuned by minimization of the SCL loss (Khosla et al., 2020) to push together the embeddings of the word occurrences that have the same supersense label while pushing apart the embeddings of the word occurrences with different supersense labels.[7] See Figure 1, (b), for an illustration. The SCL loss for a batch is given by Formula 1, where f is a vector similarity metric, τ is the temperature parameter[8], I is the set of anchors in the batch (all noun and verb occurrences), for each $i \in I$, $P(i)$ is the set of positive instances for i

[4]We experimented with freezing either no layer, the first 4 layers, the first 8 layers, or the first 11 layers. Freezing the first 8 layers gave the best WiC task performance.

[5]Code is available at: `https://github.com/yavasde/Fine-Tuning-For-Sense-Enhancement/`

[6]As these LMs works with subword tokens, we consider two labels for each supersense; one (`B-`) used for tokens starting a word occurrence, and one (`I-`) used for other tokens.

[7]We use as embedding of a word occurrence the average of the embeddings of its subword tokens, as suggested by Mosolova et al. (2024).

[8]The temperature controls the separation between positive and negative instances. See Section 5 for details.

(i.e., word occurrences with the same label as i, excluding i itself), and $A(i)$ is the set of all positive or negative instances for i in the batch (again, excluding i itself).

$$\mathcal{L}_{\text{SCL}} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \left(\frac{\exp\left(\frac{f(\mathbf{z}_i, \mathbf{z}_p)}{\tau}\right)}{\sum_{a \in A(i)} \exp\left(\frac{f(\mathbf{z}_i, \mathbf{z}_a)}{\tau}\right)} \right) \quad (1)$$

The batches are sampled randomly. Following previous studies (Chen et al., 2020; Gao et al., 2021; Yan et al., 2021; Mosolova et al., 2024), we use *Cosine Similarity*.

We experiment with different temperature values for both SCL and SPL in Section 5. The details about the hyperparameters and the training can be found in Appendix A.

3.3 Fine-Tuning for Task Adaptation

We experiment with fine-tuning the LMs with an objective similar to the downstream task (WiC). Specifically, we fine-tune the LMs via SPL to classify word occurrence pairs and determine if they share the same supersense. The goal is to investigate how task similarity affects downstream task performance. More precisely, we aim to investigate how adjusting the embedding space to suit the downstream task compares to enhancing it with sense knowledge in terms of performance.

3.3.1 Data Preprocessing

Starting from SemCor as outlined in Section 3.1, we randomly pair sentences containing the same word (either verb or noun) and label the pair depending on whether this word shares the same supersense in both sentences. We only consider words that occur in the corpus with at least two supersenses.

3.3.2 Method

We concatenate the sentence pairs with a separator token in between and feed the result to the LM. We concatenate the embeddings of the two occurrences of the target word (produced by the final layer of the model) and feed the result to a binary classification layer.[9] The LM is fine-tuned via SPL using Cross-Entropy Loss. Details about the hyperparameters and the training can be found in Appendix A.

3.4 Isotropization

We apply the isotropization post-processing method proposed by Mu and Viswanath (2018) to increase the isotropy of the original BERT embeddings. This post-processing method is applied to the embeddings of the test set, which are obtained as described in Section 4.1, to evaluate their topological properties, and also to the WiC dataset to assess task performance with isotropized embeddings. The goal is to evaluate the effects of isotropization in isolation on task performance, without adding additional knowledge.

The method involves three steps. First, the embeddings are centered by computing their mean vector and subtracting the mean vector from each embedding. Then, Principal Component Analysis is applied to the centered embeddings to identify the components that carry high variance. Finally, for some $k \in \mathbb{N}$ (in our case, $k = 1$), the k top principal components are removed from each centered embeddings — this is done by subtraction of its projections on these k directions.[10] Because the principal components are the directions with the highest variance, removing these components helps create a more uniform distribution of directions.

4 Evaluation of the Embeddings

We evaluate the embeddings from both BERT and RoBERTa, including the original embeddings (with and without isotropization), embeddings of the models fine-tuned for sense knowledge enhancement using SCL and SPL, and embeddings of the models fine-tuned for task adaptation. The evaluation is based on both the topology of the embedding space and the performance on the WiC task.[11]

4.1 Topology of the Embedding Space

We evaluate the quality of the embeddings based on their topological properties using three metrics: *sense-alignment*, *uniformity*, and *isotropy*. This evaluation is performed on the test set (see Section 3.1), using nouns and verbs only. The embeddings are extracted from the final layer of the LM and are L2-normalized before the evaluation. We write $E = [e_1, \ldots, e_N]$ the matrix of all extracted embeddings.

Sense-Alignment: This metrics is defined as the average over all supersenses of the average

[9] We use as embedding of a word occurrence the average of the embeddings of its subword tokens.

[10] Our experiments show that removing the first top principal component is enough to substantially increase uniformity and isotropy. For detailed experiments, see Appendix B.

[11] We conducted additional experiments with the Word Sense Disambiguation task (Raganato et al., 2017) and observed performance patterns similar to those seen in the WiC task across different embeddings. Details of these experiments can be found in Appendix C.

	Topological Properties			WiC	
	Uniformity	Sense-Alignment	Isotropy	Threshold	Classifier
BERT	2.84	0.33	0.58	63.2	55.9 (0.7)
+*Isotropization*	**3.95**	0.04	**0.98**	**63.6**	**56.7** (0.2)
+*SCL*	2.81	**0.72**	**0.92**	**64.7**	**56.8** (0.4)
+*SPL*	2.49	**0.52**	0.55	**64.0**	55.8 (0.5)
+*Task Adaptation*	**3.37**	0.22	**0.68**	62.9	**56.5** (0.5)
RoBERTa	0.73	0.83	0.40	61.2	56.5 (0.5)
+*Isotropization*	**3.89**	0.07	**0.97**	**63.2**	**56.7** (0.5)
+*SCL*	**2.92**	0.71	**0.90**	**64.8**	**59.3** (0.3)
+*SPL*	**2.95**	0.67	**0.70**	**65.2**	**58.8** (0.6)
+*Task Adaptation*	**1.87**	0.55	**0.46**	61.2	56.5 (0.9)

Table 1: Evaluation results of different embeddings in terms of topological properties and task performance. Higher values are better for all topological property metrics. Accuracy is reported for *WiC Threshold* and *WiC Classifier*. The mean results over 5 runs are given with standard deviation in brackets for WiC Classifier. The values that are better than the original LM embeddings are given in bold. The models with the best WiC task performance across various temperature settings are reported (*BERT+SCL*: 0.3 *BERT+SPL*: 0.1, *RoBERTa+SCL*: 0.2 *RoBERTa+SPL*: 4.5).

pairwise cosine similarity for word occurrences with this supersense. Higher values indicate better sense-alignment.

Uniformity: This metrics measures the density of the distribution of the embeddings in their vector space (Wang and Isola, 2020). It is defined as

$$-\log\left(\frac{2}{N(N-1)}\sum_{1\leq i<j\leq N}\exp(-2\|e_i-e_j\|_2^2)\right) \quad (2)$$

Higher values indicate more uniform distributions of the embeddings.

Isotropy: This metrics measures the distribution of the embeddings across all directions of the embedding space (Mu and Viswanath, 2018). It is defined as

$$\frac{\min_{u\in U} F(u)}{\max_{u\in U} F(u)} \quad (3)$$

where U is the set of eigenvectors of E^TE and F is the *partition function* introduced by Arora et al. (2016) as

$$F(u)=\sum_{i=1}^{N}\exp(u\cdot e_i) \quad (4)$$

The isotropy score ranges from 0 to 1 and higher values indicate more isotropic distributions.

4.2 WiC Performance

The WiC task is a binary classification task where the goal is to determine if two occurrences of a same word form, either both verbs or both nouns, have the same meaning. We evaluate the WiC performance of two methods, as proposed in the original paper (Pilehvar and Camacho-Collados, 2019): *WiC Threshold* and *WiC Classifier*. Comparing these two methods allows us to better understand how different topological properties of the embeddings help to solve this task in two setups, unsupervised and supervised.

For both methods, we extract the embeddings produced by the last 4 layers of the LM and average them to obtain a single embedding per token, as done in previous work (Liu et al., 2021b). Additionally, in cases where the target word is tokenized into subwords by the model tokenizer, we average all subword embeddings to obtain one embedding per word.

WiC Threshold: Two occurrences of a word are classified as having the same meaning if the cosine similarity of their embeddings is higher than a fixed threshold. This threshold is tuned on the development set and we report the accuracy achieved with the best threshold on the test data.[12]

WiC Classifier: We train a binary classifier to predict whether two occurrences of a word have the same meaning. The classifier takes the concatenation of their embeddings as input. We use a feed-forward network with one hidden layer with ReLU activation.[13] We train five randomly initialized classifiers for each type of embedding and report the mean accuracy.

4.3 Results

Table 1 summarizes our results; we report the results for the best temperature for each fine-tuning method, determined by the average performance on both WiC task setups.

[12] We have experimented with various similarity metrics, including *dot product* and *Euclidean Distance*, and found that *Cosine Similarity* yields the highest accuracy scores overall.

[13] See Appendix A for details on the hyperparameters and training.

Language Model-Specific Analysis. BERT and RoBERTa embeddings exhibit different topological properties. More specifically, RoBERTa embeddings are more sense-aligned, less uniform, and less isotropic. Additionally, embeddings of two LMs differ in performance: BERT embeddings yield better performance on WiC Threshold, whereas RoBERTa embeddings yield better performance on WiC Classifier. Overall, fine-tuned RoBERTa embeddings yield better results on both setups, and the improvements achieved are more significant compared to BERT embeddings.

Comparison of Fine-Tuning Methods. For both LMs, isotropization leads to improved performance on both WiC Threshold and WiC Classifier compared to the original embeddings. SCL and SPL produce embeddings that differ in terms of topological properties, however, both methods yield embeddings with similar properties across the two LMs. Compared to SPL, SCL generates embeddings that are more isotropic and more sense-aligned with both models. Uniformity values, however, varies slightly depending on the LM.

Considering WiC performance, SCL achieves the highest average WiC scores with both BERT and RoBERTa, outperforming other methods. With BERT, SPL embeddings outperform the original LM embeddings only on WiC Threshold, showing limited and inconsistent improvements in overall WiC performance. Lastly, fine-tuning for task adaptation does not lead to consistent improvements: it improves performance on WiC Classifier with BERT, but reduces it on WiC Threshold, and shows no gains for either task setup with RoBERTa.

The differences in isotropy and sense-alignment between SCL (higher) and SPL (lower) might be somewhat surprising when considering the two training objectives. SCL works by explicitly pushing closer to each other embeddings of same superpense tokens and pushing farther embeddings of different supersense tokens. SPL, when implemented with a linear classification layer, works in a similar way, although indirectly: each supersense is represented by a vector in the embedding space, and token embeddings are pushed both closer to their supersense vector and farther from other supersense vectors. One obvious difference between the two methods is that we interpret *closer/farther* in terms cosine similarity for SCL and in terms of dot product for SPL. Another one is the use of class vectors in SPL which, depending on their random initialization, might cause SPL to affect the embedding space in more arbitrary ways.

Topological Properties and Task Performance. Our results are in line with previous studies that show the importance of isotropy in dowstream task performance (Biś et al., 2021; Gao et al., 2021; Liu et al., 2021a,b; Rajaee and Pilehvar, 2021a). We obtain higher WiC performance with more isotropic embeddings. However, the best performances are achieved with the embeddings that are both more isotropic and sense-aligned with both LMs.

Additionally, high sense-alignment alone does not guarantee strong task performance. RoBERTa embeddings exhibit lower WiC Threshold scores compared to BERT embeddings despite being more sense-aligned. Similarly, fine-tuning BERT with SPL increases sense-alignment without increasing isotropy or uniformity. This only improves performance on WiC Threshold while reducing performance on WiC Classifier. These findings are consistent with previous studies emphasizing the importance of balancing multiple metrics (Wang and Isola, 2020; Wang and Liu, 2021; Gao et al., 2021).

In contrast to these studies, we observes that increased uniformity does not always yield better results and that its effects vary by model. With BERT, the best-performing embeddings have similar (but slightly lower) uniformity values as the original embeddings. With RoBERTa, however, increasing uniformity and isotropy leads to better results. Note that these LMs differ notably in terms of their initial sense-alignment and uniformity values.

Lower uniformity indicates that embeddings are more concentrated in specific regions of the space. In our case, these clusters likely arise from sense-alignment, where the regions are influenced by sense similarities.[14] However, the results show that maintaining the distribution of embeddings (uniformity) as much as possible while increasing sense similarities is more beneficial. This suggests a trade-off between these metrics, and balancing them proves more beneficial. See Appendix D for the visualization of the embedding spaces, illustrating the relation between sense-alignment and uniformity.

[14]The studies mentioned evaluate the impact of uniformity in other contexts. When embeddings are intentionally aligned with specific knowledge, reduced uniformity may arise from meaningful similarities, whereas, e.g., in encoder training, it may arise from unwanted artifacts in the data.

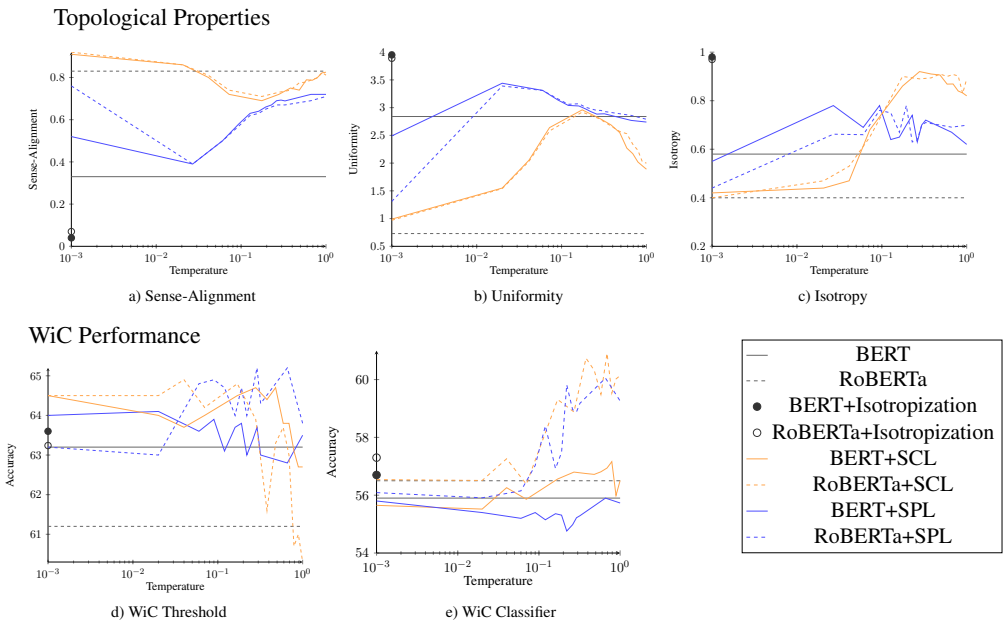

Figure 2: Relation between temperature, different topological properties, and WiC performance. We report the results for the embeddings of LMs fine-tuned with SCL and SPL with different temperature values (normalized), the original LM embeddings, and isotropized LM embeddings (indicated by markers). Three topological properties are studied: *sense-alignment* (a), *uniformity* (b), and *isotropy* (c). For WiC task performance, *WiC Threshold* (d) and *WiC Classifier* (e) performances are studied. The mean results over 5 runs are given for WiC Classifier.

5 Relation Between Temperature and Topology of the Embedding Space

In SCL, the temperature parameter of the loss controls the strength of the separation between positive and negative instances. A lower temperature increases the influence of harder-to-separate instances, effectively generating harder negatives. As for SPL, the temperature parameter scales the probability distribution of the classes and higher values of temperature lead to softer probability distributions.

We experiment with different temperature values to investigate the effects of temperature on the topology of the embedding space. Furthermore, we show how the embeddings with different topological properties perform differently on the WiC task. These experiments provide a more in-depth understanding of the fine-tuning methods and gives us a more complete picture of how variations in topological properties relate to task performance.

Temperature is already a parameter in the SCL loss function (see Formula 1). We introduce a temperature parameter into the Cross-Entropy loss by scaling the logits before applying the softmax. In the context of SCL loss, the temperature typically ranges between 0 and 1, whereas in Cross-Entropy loss, the temperature can vary over a wider range, typically between 0.1 and 100 (Agarwala et al., 2020). We experiment with 13 temperature values for both loss functions ranging between 0.03 and 1 for SCL and between 0.1 and 15 for SPL.[15]

5.1 Results

See Figure 2 for the relation between temperature, topological properties, and WiC performance.[16] Overall, temperature affects the topology of the embedding space. A clear trade-off is observed between sense-alignment, isotropy, and uniformity. As sense-alignment decreases, both uniformity and isotropy tend to increase.

Comparison of Fine-Tuning Methods. Different fine-tuning methods create embeddings with different properties. Overall, SCL creates more sense-aligned, more isotropic, but less uniform embeddings. On the other hand, SPL creates less sense-aligned, less isotropic, but more uniform embeddings. Furthermore, both methods produce embeddings with similar topological properties with both LMs, despite the initial topological differences between the models.

[15] The selected temperatures for SCL are: 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. The selected temperatures for SPL are: 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 10, 15.

[16] See Appendix E for complete values.

Considering task performance, both fine-tuning methods lead to performance improvement with RoBERTa, while only SCL leads to performance improvement with BERT. With BERT, SPL embeddings do not outperform the original BERT embeddings at any temperature value on WiC Classifier and they only outperform them in early temperatures on WiC Threshold. Overall, SCL yields better results, demonstrating high performance with both models and task setups.

Language Model-Specific Analysis. The performance trends are similar across models, however, the improvements with RoBERTa are more substantial, despite both LMs having embeddings with similar properties after fine-tuning. With RoBERTa, both fine-tuning methods yield more comparable performance, whereas with BERT, SCL shows a clear performance advantage.[17]

Unsupervised and Supervised Tasks. Each task exhibits different trends. On WiC Threshold, the fine-tuned embeddings generally outperform the original LM embeddings. Sense-alignment seems to play a significant role after isotropy reaches a certain threshold. Notably, less isotropic but more sense-aligned BERT+SPL embeddings outperform isotropized BERT embeddings. Overall, high performances are achieved when both isotropy and sense-alignment exceed certain levels.

WiC Classifier performance varies more across models and fine-tuned model embeddings do not always surpass the original LM embeddings in terms of performance. Both high levels of isotropy and sense-alignment seem important. For example, with BERT, surpassing the isotropized embeddings is challenging and only happens at the latest temperatures with SCL, when isotropy and sense-alignment are both significantly high. Similarly, highly sense-aligned original RoBERTa embeddings already start with better performance than the BERT embeddings and the best performances are achieved with RoBERTa when both properties are significantly high.

6 Conclusion

In this study, we explore the relation between different fine-tuning methods, the topology of the resulting embedding spaces and the downstream task performance in the context of sense knowledge enhancement with two pre-trained language models (LMs), BERT and RoBERTa. We focus on three topological properties: *sense-alignment*, *uniformity*, and *isotropy*.

Our comparison of two fine-tuning methods, *Supervised Contrastive Learning (SCL)* and *Supervised Predictive Learning (SPL)*, shows that these methods create embedding spaces that differ in terms of topological properties and the resulting embeddings achieve different task performance. Although the embeddings of BERT and RoBERTa differ in their initial properties, fine-tuning methods affect the embedding spaces of both LMs in similar ways, resulting in embeddings with similar properties. Overall, SPL — the standard fine-tuning method — shows mixed results, with effectiveness varying by LM. It produces embeddings that outperform the original LM embeddings in most cases, though not consistently. In contrast, SCL demonstrates more consistent improvements with both LMs and generally outperforms SPL.

With respect to topological properties, our results show that the embeddings with increased isotropy achieve better task performance. However, the best-performing embeddings are those with high sense-alignment and isotropy. Embeddings with only high sense-alignment do not always yield better results and can underperform compared to the original LM embeddings. Furthermore, we observe a trade-off between uniformity and sense-alignment, with optimal results achieved by balancing the two. Overall, our results highlight the importance of balancing all three properties for optimal task performance.

Moreover, our results show that the degree to which these topological properties are beneficial in a downstream task differs between a supervised and an unsupervised system. Both sense-alignment and isotropy contribute to performance, but their relative importance varies across task setups. Finally, our findings suggest that fine-tuning an LM on a task similar to the downstream task does not necessarily produce embeddings that improve downstream task performance.

In conclusion, our findings highlight the importance of topological properties of the embeddings during knowledge enhancement, demonstrating that fine-tuning an LM without considering these properties may not always lead to improved downstream task performance.

[17] The difference in improvement between the LMs may be influenced by factors beyond the scope of our current analysis, warranting further investigation.

Limitations

While the results of this study offer valuable insights, several limitations must be acknowledged. The main limitation of this study is that our experiments only focus on English, which limits the generalization of our findings to other languages. However, we expect similar results with other languages as topological properties of the embeddings are more closely linked to factors such as learning objectives, loss functions used for the training of the models.

Similarly, our experiments rely on WordNet supersenses and the SemCor corpus, which limits the generalization of our results to other sense inventories or datasets. However, these resources are widely used and highly valuable for word sense tasks.

We further acknowledge that there may be other properties influencing the success of language model embeddings, and our study focuses on only three specific properties. However, isotropy and uniformity are commonly used metrics, especially in terms of expressiveness. Furthermore, our study offers potential explanations for the observed performance differences across different types of embeddings with these properties.

Finally, we rely on Cosine Similarity as a similarity metric for both training and evaluation of the models. However, this metric is the most widely used and successful semantic similarity metric. Using this metric in both stages does not influence the core findings of our study; instead, it highlights the importance of isotropy.

Acknowledgement

This study is funded by the DFG project “Coercion and Copredication as Flexible Frame Composition". We would like to thank the anonymous reviewers for their valuable comments.

References

Atish Agarwala, Jeffrey Pennington, Yann Dauphin, and Sam Schoenholz. 2020. Temperature check: theory and practice for training models with softmax-cross-entropy losses. *arXiv preprint arXiv:2010.07344*.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49, Barcelona (online). International Committee for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Geetanjali Bihani and Julia Rayz. 2021. Low anisotropy sense retrofitting (LASeR) : Towards isotropic and sense enriched representations. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 81–95, Online. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too much in common: Shifting of embeddings in transformer language models and its implications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5117–5130, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

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, Hong Kong, China. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2020. MULTISEM at SemEval-2020 task 3: Fine-tuning BERT for lexical meaning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 158–165, Barcelona (online). International Committee for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *International conference on machine learning*.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021a. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021b. MirrorWiC: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Anna Mosolova, Marie Candito, and Carlos Ramisch. 2024. Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 34–41, St. Julian's, Malta. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of*

the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021a. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021b. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Younes Samih and Laura Kallmeyer. 2023. Unsupervised semantic frame induction revisited. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 89–93, Nancy, France. Association for Computational Linguistics.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019a. Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pages 6555–6565. PMLR.

Feng Wang and Huaping Liu. 2021. Understanding the Behaviour of Contrastive Loss . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, Los Alamitos, CA, USA. IEEE Computer Society.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2019b. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Deniz Ekin Yavas. 2024. Assessing the significance of encoded information in contextualized representations to word sense disambiguation. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 42–53, Malta. Association for Computational Linguistics.

Deniz Ekin Yavas, Timothée Bernard, Laura Kallmeyer, and Benoît Crabbé. 2024. Improving word sense induction through adversarial forgetting of morphosyntactic information. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 238–251, Mexico City, Mexico. Association for Computational Linguistics.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 532–542, Seattle, United States. Association for Computational Linguistics.

Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527, Online. Association for Computational Linguistics.

A Model Hyperparameters and Training Details

The training of the models is done on NVIDIA RTX A5000 GPU, with a single GPU. The PyTorch library (Paszke et al., 2019) is utilized for the training. We use BERT (*base-uncased*) for English with parameter size 110M and RoBERTa (*base*) for English with parameter size 125M, which are accessed via the Transformers library (Wolf et al., 2020). The SemCor corpus, which is the data used for fine-tuning, is accessed via the NLTK Library (Bird and Loper, 2004). Finally, the WiC

Dataset, evaluation data, is accessed via https://super.gluebenchmark.com/tasks. Both SemCor and WiC are created for the English language.

Licenses for the model and data used:

- WiC: Creative Commons Attribution-NonCommercial 4.0 License.
- BERT: Apache license 2.0

A.1 Fine-Tuning for Sense Knowledge Enhancement

A.1.1 Supervised Contrastive Learning

- **Optimizer:** Adam
- **Batch Size:** 8
- **Learning Rate:**
  - BERT: $1e-05$,
  - RoBERTa: $1e-05$, $1e-04$ after τ=0.5,
- **Number of Epochs (best τ):**
  - BERT: 4,
  - RoBERTa: 4,
- **τ (best):**
  - BERT: 0.3,
  - RoBERTa: 0.2,
- **Loss:** SCL
- We employ early stopping based on validation loss.

A.1.2 Supervised Predictive Learning

- **Classification Head:** A dense layer with a Softmax activation function
- **Optimizer:** Adam
- **Batch Size:** 8
- **Learning Rate:**
  - BERT: $1e-05$,
  - RoBERTa: $1e-05$, $1e-04$ after τ=4,
- **Number of Epochs (best τ):**
  - BERT: 3,
  - RoBERTa: 2,
- **τ (best):**
  - BERT: 0.1,
  - RoBERTa: 4.5,
- **Loss:** Cross-Entropy
- We employ early stopping based on validation loss.

A.2 Fine-Tuning for Task Adaptation

- **Classification Head:** A dense layer with a Sigmoid activation function
- **Optimizer:** Adam
- **Batch Size:** 8
- **Learning Rate:**
  - BERT: $5e-06$,
  - RoBERTa: $1e-06$
- **Number of Epochs:**
  - BERT: 2,
  - RoBERTa: 4,
- **Loss:** Cross-Entropy
- We employ early stopping based on validation loss.

A.3 WiC Task

- A feed-forward network with one hidden layer with ReLU activation and an output layer with Sigmoid activation.
- **Optimizer:** Adam
- **Batch Size:** 32
- **Learning Rate:** $1e-04$
- **Drop-out:** 0.2
- We employ early stopping based on validation loss.
- We train a classifier five times using each type of embedding and report the mean test accuracy.

B Isotropization Experiments

For the isotropization post-processing method, we test the effect of removing various numbers of top components from the BERT embeddings, specifically 1, 5, 10, 15, and 20 components. The quality of the resulting embeddings is presented in Table 2.

This method successfully increases the isotropy and uniformity of BERT embeddings. The number of top components removed influences the trade-off between isotropy and the retention of semantic information in the embeddings. Isotropy and uniformity improve after removing just the first top component, with both continuing to improve as more components are removed. Furthermore, we notice that sense-alignment drops significantly only after the first top component is removed. As more components are removed, less semantic information is preserved, and sense-alignment continues to reduce.

C Word Sense Disambiguation Task

The Word Sense Disambiguation (WSD) task (Raganato et al., 2017) involves selecting the correct WordNet sense label for a target word in context from its candidate senses. Similar to our approach,

Model Type	#	Uniformity	Sense-Alignment	Isotropy
BERT	-	2.840	0.332	0.588
BERT+Isotropization				
	1	3.958	0.045	0.986
	5	3.967	0.031	0.988
	10	3.971	0.022	0.993
	15	3.972	0.018	0.995
	20	3.973	0.015	0.995

Table 2: Topological properties of the BERT embeddings and isotropized BERT embeddings with different numbers of components removed. We report 3 topological properties: *sense-alignment*, *uniformity*, and *isotropy*. Higher values are better for all topological properties.

Figure 3: The t-SNE visualizations of the test data using embeddings from different models. Models: BERT embeddings (left), BERT+SCL embeddings with $\tau = 0.2$ (middle), BERT+SCL embeddings with $\tau = 1.0$ (right). Different colors refer to different supersenses.

Model	WSD
BERT	48.1
+*Isotropization*	47.9
+*SCL*	**48.9**
+*SPL*	**48.8**
+*Task Adaptation*	**48.7**
RoBERTa	49.0
+*Isotropization*	**50.4**
+*SCL*	**53.3**
+*SPL*	**52.2**
+*Task Adaptation*	**50.6**

Table 3: WSD results using different embeddings. *F1* is reported. The values that are better than the original LM embeddings are given in bold. The results of the models with the best WiC task performance across various temperature settings are reported (*BERT+SCL*: 0.3 *BERT+SPL*: 0.1, *RoBERTa+SCL*: 0.2 *RoBERTa+SPL*: 4.5).

SemCor is used as the training corpus in the original task. Following previous work (Liu et al., 2021b), we design a one-shot setting in which each sense is represented by a single example from WordNet. We compute the cosine similarity between the embedding of the target word in the example sentences and that of the test instance, and select the sense label with the highest similarity. We exclude the senses that do not have any examples in WordNet.

The results with the embeddings of the LMs and fine-tuned LMs (best temperature) can be seen in Table 3. While isotropization does not always improve the performance, the SCL fine-tuning method achieves the best results with both LMs.

D Visualization of the Embedding Spaces

Since uniformity is measured using the pairwise similarities between embeddings, the decrease in uniformity observed in sense-enhanced embeddings may be associated with the formation of dense sense clusters. As sense-alignment increases and uniformity decreases, instances of the same sense may become closer. This relation can be captured with t-SNE visualization, as it is useful for capturing the local relations in the data.

The t-SNE visualizations of the test set using embeddings from different models are shown in

Figure 3. The models and the topological properties of their embeddings are as follows:

- **BERT:**
  Sense-Alignment: 0.33, Uniformity: 2.84
- **BERT+SCL with $\tau = 0.2$:**
  Sense-Alignment: 0.72, Uniformity: 2.96
- **BERT+SCL with $\tau = 1.0$:**
  Sense-Alignment: 0.82, Uniformity: 1.89

We observe that the embeddings of BERT+SCL ($\tau = 1.0$) with reduced uniformity show to form more dense clusters of same sense instances.

E Temperature Details

The complete topological property and WiC performance values for embeddings extracted from models with different temperatures are provided in Table 4.

Model Type	Temperature	Topological Properties			WiC Performance	
		Uniformity	Sense-Alignment	Isotropy	Threshold	Classifier
BERT	-	-2.83	0.33	0.58	63.2	55.90 (0.7)
BERT+SPL						
	0.1	2.49	0.52	0.55	64.0	55.8 (0.5)
	0.5	3.44	0.39	0.78	64.1	55.4 (0.3)
	1	3.31	0.50	0.69	63.6	55.28 (0.8)
	1.5	3.14	0.58	0.78	63.9	55.43 (0.5)
	2	3.04	0.63	0.64	63.1	55.15 (0.1)
	2.5	3.03	0.64	0.65	63.7	55.36 (0.8)
	3	2.98	0.66	0.70	63.8	55.31 (0.2)
	3.5	2.93	0.67	0.74	63.0	54.76 (0.4)
	4	2.88	0.69	0.63	63.4	54.98 (0.4)
	4.5	2.88	0.69	0.70	63.7	55.22 (0.7)
	5	2.88	0.68	0.72	63.0	55.29 (0.6)
	10	2.77	0.72	0.67	62.8	55.9 (0.8)
	15	2.74	0.72	0.62	63.5	55.73 (0.4)
BERT+SCL						
	0.03	0.99	0.91	0.42	64.5	55.65 (0.3)
	0.05	1.55	0.86	0.44	64.0	55.52 (0.6)
	0.07	2.05	0.80	0.47	63.7	56.26 (0.4)
	0.1	2.64	0.72	0.68	64.0	55.86 (1.1)
	0.2	2.96	0.69	0.86	64.5	56.56 (0.5)
	0.3	2.81	0.72	0.92	64.7	56.8 (0.4)
	0.4	2.69	0.75	0.91	64.4	56.75 (0.4)
	0.5	2.55	0.74	0.90	64.7	56.72 (0.3)
	0.6	2.27	0.78	0.86	63.8	56.81 (0.6)
	0.7	2.17	0.78	0.86	63.8	56.94 (0.7)
	0.8	2.02	0.80	0.84	63.1	57.16 (0.7)
	0.9	1.95	0.82	0.83	62.7	55.99 (0.4)
	1	1.89	0.82	0.82	62.7	56.51 (0.3)
RoBERTa	-	0.73	0.83	0.40	61.2	56.50 (0.5)
RoBERTa+SPL						
	0.1	1.31	0.76	0.44	63.2	56.09 (0.4)
	0.5	3.39	0.39	0.66	63.0	55.91 (0.6)
	1	3.31	0.50	0.66	64.8	56.15 (0.5)
	1.5	3.16	0.57	0.76	64.9	57.03 (0.4)
	2	3.06	0.62	0.75	64.7	58.40 (0.6)
	2.5	3.07	0.63	0.67	64.0	56.92 (0.5)
	3	3.01	0.65	0.78	64.7	57.48 (0.9)
	3.5	2.98	0.66	0.63	64.0	59.82 (0.2)
	4	2.95	0.67	0.63	64.7	58.96 (0.8)
	4.5	2.95	0.67	0.70	**65.2**	58.88 (0.6)
	5	2.93	0.67	0.71	64.3	59.18 (0.6)
	10	2.86	0.69	0.69	**65.2**	**60.05** (1.3)
	15	2.79	0.71	0.69	63.8	59.26 (0.8)
RoBERTa+SCL						
	0.03	0.97	0.92	0.40	64.5	56.54 (0.3)
	0.05	1.54	0.86	0.47	64.5	56.51 (0.3)
	0.07	2.03	0.81	0.53	**64.9**	57.26 (0.4)
	0.1	2.58	0.74	0.65	64.2	56.38 (0.3)
	0.2	2.92	0.71	0.90	64.8	59.31 (0.3)
	0.3	2.80	0.73	0.89	63.9	58.88 (0.6)
	0.4	2.67	0.74	0.89	61.6	60.74 (0.3)
	0.5	2.57	0.77	0.91	63.3	60.32 (0.1)
	0.6	2.53	0.77	0.90	63.7	59.49 (0.2)
	0.7	2.34	0.79	0.90	63.0	**60.91** (0.3)
	0.8	2.23	0.79	0.90	60.7	59.53 (0.3)
	0.9	2.02	0.82	0.83	61.0	60.02 (0.7)
	1	2.00	0.81	0.89	60.3	60.14 (0.4)

Table 4: Topological properties and WiC results of the embeddings of different models with different temperature parameter values. The results for original LM embeddings (BERT and RoBERTa) are given for comparison. Accuracy is given for *WiC Threshold* and *WiC Classifier* results. For WiC Classifier, the mean results over 5 runs are given with standard deviation in brackets. Higher values are better for all topological property metrics.