Is Anisotropy Really the Cause of BERT Embeddings not Being Semantic?

Alejandro Fuster-Baggetto
Universidad Nacional de
Educación a Distancia (UNED)
Voicemod S.L.
afuster53@alumno.uned.es
alejandrofuster1@gmail.com

Víctor Fresno
Universidad Nacional de
Educación a Distancia (UNED)
vfresno@lsi.uned.es

Abstract

In this paper we conduct a set of experiments aimed to improve our understanding of the lack of semantic isometry in BERT, i.e. the lack of correspondence between the embedding and meaning spaces of its contextualized word representations. Our empirical results show that, contrary to popular belief, the anisotropy is not the root cause of the poor performance of these contextual models' embeddings in semantic tasks. What does affect both the anisotropy and semantic isometry is a set of known biases: frequency, subword, punctuation, and case. For each one of them, we measure its magnitude and the effect of its removal, showing that these biases contribute but do not completely explain the phenomenon of anisotropy and lack of semantic isometry of these contextual language models.

1 Introduction

The introduction of the Transformer architecture (Vaswani et al., 2017) has had an enormous impact over Natural Language Processing (NLP) in the last years. These models, trained for language modelling via self-supervision over huge text collections, have become the state-of-the-art in many NLP tasks by applying a fine-tuning process with a small supervised dataset.

Semantic Textual Similarity (STS) is one of these tasks, where training a linear layer on top of BERT (Devlin et al., 2019) to get the similarity score of two concatenated sequences has emerged as a new standard in this field. However, this approach, called cross-encoder, has its own problems. It takes two sentences as input, so it has to rerun the whole model for each pair of sentences. This makes infeasible the application of this approach to most tasks.

In order to solve this problem of efficiency, a lightweight approach called bi-encoder, consisting on obtaining the distance of two sentences within the embedding space, has emerged. In this case, each sentence is passed separately to the Transformer, which obtains an embedding for each of them. A similarity metric is then computed between the two embeddings. Unfortunately, this approach does not perform well with vanilla pre-trained Transformers. Reimers and Gurevych (2019) showed that representing sentences by average contextual BERT embeddings perform worse than averaging static Glove embeddings in semantic tasks (Pennington et al., 2014), despite their lack of contextuality.

On the other hand, studying this mismatch between the contextual word embedding and meaning spaces, Gao et al. (2019) diagnosed high anisotropy in Transformer language models, named as the *representation degeneration problem*. This means that embeddings do not follow a uniform distribution with respect to direction, i.e. the embeddings concentrate in an hypercone instead of occupying the whole space. They stated that this degeneration of the representation space could be closely related with the lack of semantic isometry to the point that some authors have tried to correct it by applying isotropy correction techniques (Li et al., 2020; Su et al., 2021). On the other hand, approaches based on contrastive learning have recently achieved remarkable improvements in this respect (Zhang et al., 2021; Giorgi et al., 2021; Yan et al., 2021; Gao et al., 2021).

Contrastive learning methods pull together semantically similar sentence vectors and push apart semantically dissimilar ones, partially correcting the semantic isometry of embedding spaces. Despite their good results, these techniques also have their own problems. The main difference between these methods is how they perform the selection of the positive (similar) and negative (dissimilar) sentence pairs they need for training.

For example, Zhang et al. (2021) used back translation from English to German to obtain augmented

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4271–4281

views of a sentence, and Giorgi et al. (2021) used near text spans in a document as positive samples. Finally, Gao et al. (2021) used dropout as a data augmentation technique for the unsupervised model, and the NLI dataset (Bowman et al., 2015) annotations for the supervised model. Despite their originality, all of these approaches have different weaknesses, and they are not easily improved, as authors just find better ways of creating positive and negative pairs, without really building on previous work. Consequently, improvements in semantic isometry correctness obtained by one method do not serve as a starting point for the next method, and no progress is made in solving the overall problem.

Contrastive bi-encoder models achieve unprecedented results in semantic tasks, but the root cause of the anisotropy and lack of semantic isometry observed in pre-trained Transformer language models is not fully understood. We think that any finding in this area can be very relevant and open new research lines that can lead to future improvements in bi-encoder methods.

Therefore, our main aim in this research is the improvement of our understanding of the BERT embedding space, and the relationship between semantics and isotropy through empirical results. After our experimentation, we conclude that there is not enough evidence to say that anisotropy is the root cause of the lack of semantic isometry of BERT embeddings, while some biases seem to affect both isotropy and semantic isometry. We name bias to any information from a sentence that is encoded in the embedding space and that is not relevant to its meaning.

This paper is structured as follows. In Section 2 we review the related work. Next, in Section 3 we describe the experiments carried out and present their corresponding analysis of results. In Section 4 we draw our conclusions, derived from our empirical results, and leave some ideas for future work. Finally, we state the limitations of this work.

2 Related work

The anisotropy found by Gao et al. (2019) when training a model for Natural Language Generation tasks through likelihood maximization is produced by the combination of the Zipfian nature of natural language and the log-likelihood loss function. This representation degradation makes the most frequent tokens to concentrate in a hypercone in the embedding representation space, having a more sparse space for infrequent tokens.

Contrastive learning methods seem to correct anisotropy to some extent as well, in addition to semantic isometry, which could make us think that the anisotropy was, in fact, part of the problem. However, (Jiang et al., 2022) realized through a series of experiments that there exist certain biases in the BERT model, and that high anisotropy is not always equivalent to poor semantic isometry. Although the insights by Jiang et al. (2022) are certainly interesting, they seem to contradict to some degree previous works like (Gao et al., 2019; Ethayarajh, 2019; Li et al., 2020).

On the other hand, Luo et al. (2021) and Kovaleva et al. (2021) had found that a big portion of the anisotropy of BERT comes from outlier dimensions, related with positional information.

Cai et al. (2020) showed that, in spite of BERT embeddings having global anisotropy, each cluster in the embedding space is isotropic, and that this local isotropy could be enough for Transformer models to achieve their full representation power. This hypothesis is supported by recent empirical results from (Ding et al., 2022). If the anisotropy comes from the existence of different clusters, and these clusters encode non-semantic information like token frequency, this can be matched with the biases described by Jiang et al. (2022) and the representation degeneration by Gao et al. (2019).

As it can be observed, there is no consensus in the literature regarding the cause of the poor performance of Transformer embeddings in semantic tasks, and there is also no consensus about the reasons for the anisotropy observed in these embedding spaces. This disagreement is amplified by the fact that there is not a standard method for evaluating anisotropy; for example, Ethayarajh (2019) evaluates it at the word level, while Jiang et al. (2022) consider the sentence level, by averaging the word embeddings. We therefore find that there is room for research and reflection on these aspects.

3 Experimentation and results

We initially carried out an exploratory analysis to better understand the magnitude of the different biases studied. Next, we compared several Transformer-based language model configurations, particularly: different pooling strategies and models, and a bias removal technique; all in terms of isotropy and semantic isometry.

A variety of studies like (Gao et al., 2019; Ethayarajh, 2019; Kovaleva et al., 2021) have shown that different Transformer-based language models have similar behaviours regarding high isotropy and poor semantic isometry, even when they differ in number of parameters, architecture or learning objective.

We considered to evaluate the following models that share the same BERT architecture: *BERT-base-uncased*, *BERT-base-cased*, *unsupervised-SIMCSE-base*, and *supervised-SIMCSE-base*. The conclusions we extract for BERT can be extended to other models, as a variety of studies like (Gao et al., 2019; Ethayarajh, 2019; Kovaleva et al., 2021) have shown that different Transformer-based language models have similar behaviours regarding high isotropy and poor semantic isometry, even when they differ in number of parameters, architecture or learning objective.

The cased and uncased models were included in order to study the case bias reported by (Jiang et al., 2022). We also include both supervised and unsupervised variants of SIMCSE (Gao et al., 2021) because it is a contrastive learning bi-encoder model that achieves state-of-the-art results in semantic tasks and it is interesting to see how much does a successful model actually increase isotropy to understand to what extent is the anisotropy related with the poor semantic performance of the non fine-tuned bi-encoders.

3.1 Exploratory analysis of biases

We call bias to any information from a sentence that is encoded in the embedding space and that is not relevant to its meaning. Four kinds of biases are defined by Jiang et al. (2022): *frequency*, *case*, *subword* and *punctuation*. All of them could partially overlap with token frequency, e.g. lowercase tokens are more frequent than uppercase ones, some punctuation marks like '.', are more common than normal words, and some subwords like '#s' (from plural) are very frequent as well, so everything could come down to the explanation given by Gao et al. (2019), of very frequent tokens being grouped in a hypercone.

To gain insight into the severity of these biases in semantic (SIMCSE) and non-semantic (BERT) embedding spaces, we sampled 1000 random sentences from the Wikipedia corpus and plotted the distributions of the similarities between pairs of embeddings of different kind of tokens. We used cosine as the similarity metric and selected the last layer word embeddings from BERT-base-uncased and unsupervised-SIMCSE-base (Figures 1, 2, and 3), except for the case bias, where we can only use BERT-base-cased (Figure 4) as there is no cased version for SIMCSE. We expected to find lower biases for SIMCSE, as its semantic isometry is higher than the non fine-tuned BERT.

3.1.1 Frequency bias

In order to evaluate the frequency bias, we first decided to compare the most frequent tokens with randomly extracted less frequent tokens; however, we finally opted to use a list of stopwords as the most frequent tokens. The reason is that the list of the most frequent tokens was mainly formed by stopwords, punctuation marks and some very common nouns like 'man' or 'woman'. We treated punctuation marks as a separate category because, despite some of them being very frequent, we did not want to assume a priori that there was not a specific bias by punctuation mark, as this would contradict the literature. On the other hand, nouns like 'man' or 'woman' are very frequent, but, as nouns, have an intrinsic meaning, unlike stopwords, that only have meaning in a syntactic context. In section 3.2 the biased tokens will be removed in order to see the effect in isotropy and semantic performance. We thought that removing nouns, even if they are frequent and biased, could be more detrimental to the meaning of the sentence than removing stopwords; for these reasons, from this point on, we will be using stopwords as a synonym of 'very frequent' tokens.

Figure 1 shows that, even in the last layer of BERT where stopword embeddings are very contextualized (Ethayarajh, 2019), the average similarity between them is still slightly higher than the similarity between stopwords and regular (less frequent) words, or directly between regular words. This confirms the frequency bias and relates it with the fact that frequent tokens are concentrated in a hypercone in the embedding space, while less frequent tokens are more sparse. It can be observed that this gap is reduced for the SIMCSE model, a more semantic model than BERT, that seems to have corrected some of this bias through contrastive learning. The contrary happens in the case of subword bias, as will be shown below.

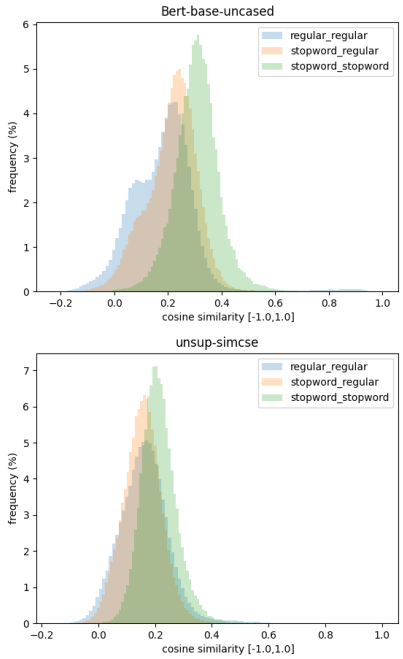

Figure 1: Average cosine similarity between stopwords and other tokens from uncased BERT (top) and unsupervised SIMCSE (bottom)

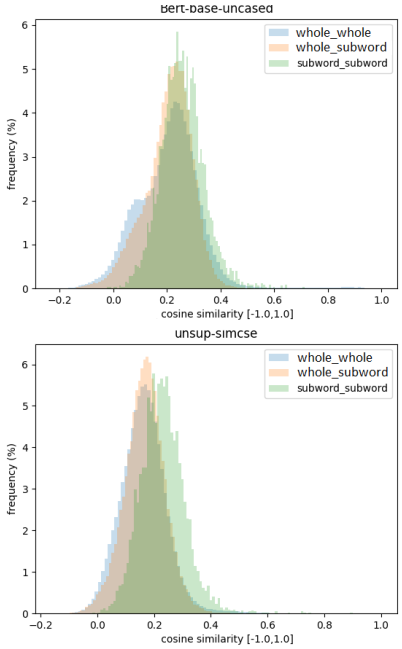

Figure 2: Cosine similarity between words and subwords from uncased BERT (top) and unsupervised SIMCSE (bottom).

3.1.2 Subword bias

Subwords are the pieces of words generated by the BERT tokenizer when it encounters Out-Of-Vocabulary (OOV) words, and their nature is varied, being Named Entities, rare words, spelling variations, or derivative and inflected words.

Figure 2 shows that the subword bias, understood as the gap of the green distribution (similarity between pairs of subwords), the blue distribution (similarity between pairs of whole words), and the orange distribution (similarity between whole words and subwords), is significantly higher in the SIMCSE model, despite its overall average cosine similarity being smaller (more isotropic) than in the BERT-base model.

3.1.3 Punctuation bias

The case of punctuation marks is more complex. For the BERT-base model, punctuation marks tend to be more sparse in average than words. Figure 3 shows the high variance of the green distribution in the BERT-base model, that represents the distance between punctuation marks. Furthermore, it also shows a cluster at the right that reaches very high similarities, being some of them near one. Although these high similarities between contextual embeddings are somewhat surprising, it all could come down to frequency.

Some punctuation marks are extremely frequent (e.g. commas or dots), others are relatively frequent (e.g. exclamations or question marks), and others are infrequent (e.g. asterisks or slashes); all of this, combined with the frequency bias discussed before, is most likely what generates the high variance of the distribution defined by cosine similarity between punctuation marks, with lower values for infrequent tokens and higher values for more frequent ones. On the other hand, SIMCSE seems to have solved the subword bias to a certain extent, although there are still relatively high similarities for some pairs of punctuation marks.

3.1.4 Case bias

Finally, the case bias is also non trivial. In the Figure 4 it can be observed that the distance between uppercase words follows a multimodal distribution, with small peaks in high similarity. This could be explained because there are two types of uppercase tokens: Named Entities, and beginning of sentence tokens. Obviously these two groups are not mutually exclusive, as a Name Entity can be at the beginning of a sentence. Both types of tokens have varied frequencies. For example, there are very common proper nouns like months and weekdays, and there are also words that are very common as sentence beginners like ‘The’. The small peaks in the highest cosine values could be due to these

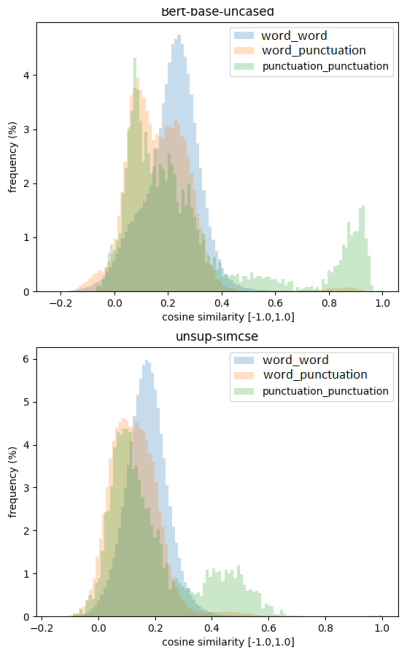

Figure 3: Cosine similarity between punctuation marks and other tokens from uncased BERT (top) and unsupervised SIMCSE (bottom)

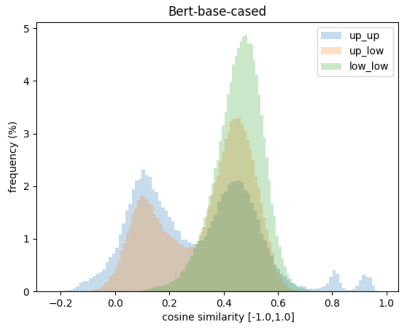

Figure 4: Cosine similarity between uppercase and lowercase words from cased BERT.

high frequency tokens. In general, the distance between lowercase words is smaller. This is expected, as these tokens are usually more frequent than uppercase ones.

3.1.5 Conclusions on bias analysis

Note that most of these points are just hypotheses that would explain the results, but that require verification. What we can conclude with certainty, though, is that, in the non fine-tuned BERT models there is indeed such things as frequency, case, and punctuation biases, while in SIMCSE we can find a certain degree of subword bias. These biases mean that a set of tokens sharing a non semantic property lay in a hypercone in the embedding space, apart from the rest of tokens that don't meet these properties, much like the clusters described by Cai et al. (2020). The mere fact that the different colors for each plot don't completely overlap and that they are sometimes multimodal supports this claim. In SIMCSE, a model with a good performance in semantic tasks, it can be observed that there is a better overlap of the different distributions. Furthermore, the distributions tend to have a lower variance and to be centered close to zero. The mean value of the distribution being around zero shows that these models are more isotropic. However, this should theoretically be irrelevant, as even if the distribution was centered in 0.8, that would only increase anisotropy and would mean that all the embeddings lay in a narrow hypercone, which, by itself, should not be problematic to semantics. What can be detrimental to the semantic performance of the model is the existence of biased clusters in the space, that would distort it with non semantic information. However, if the space is semantic and not biased, being highly anisotropic can harm the representation power of the model, but should not distort the semantic isometry.

Therefore, one of our main claims is that anisotropy is not harmful for semantics unless it is produced by a bias. This is, anisotropy is not a problem if it is the same for all tokens. If this is true, then isotropy correction techniques should not increase the performance in semantic tasks of these models, which has been empirically proven by Ding et al. (2022); Jiang et al. (2022). In the next set of experiments, we further support this idea through empirical evidence.

3.2 Isotropy vs semantic isometry

We combine and extend the experiments from (Ethayarajh, 2019; Jiang et al., 2022) regarding isotropy and semantic isometry evaluation in BERT. For isotropy evaluation, we used the previous dataset sampled from Wikipedia corpus and computed the cosine similarity between sentence embeddings in the representation space. For semantic isometry, we used the Semantic Textual Similarity Benchmark (STSB) (Cer et al., 2017) and computed the Spearman correlation between the cosine similarity of each pair of sentences and their annotation within gold standard. We conducted the following experimentation for all the layers of each model, as the literature has shown that different layers store different kinds of information.

3.2.1 Pooling comparison

First, we compared different pooling strategies for the BERT-base-uncased model. We used both the average of all the word embeddings in a sentence, and the CLS embedding as pooling strategies for obtaining sentence-level embeddings. Reimers and Gurevych (2019) pointed out that the CLS is substantially worse than the word average in semantic tasks for the non fine-tuned BERT models; however, we still thought that it was worth to include this strategy in our experimentation, especially to see how it behaved in terms of isotropy.

Additionally, we included 'none' pooling, that simply takes the word embeddings instead of pooling them into a sentence embedding, thus allowing the isotropy evaluation at the word level, like it was done by Ethayarajh (2019). Here, the cosine similarity between pairs of token embeddings is computed, instead of sentence embeddings. This can only be done for isotropy, as the benchmark for semantic isometry is only available between pairs of sentences. The results of this experiment are shown in Figure 5.

First of all, as it was previously stated, it can be observed that the semantic isometry of CLS pooling is much worse than the one of average pooling, despite being much more isotropic. On the other hand, tokens (none pooling) are not that much anisotropic in BERT-base-uncased, reaching only average similarities lower than 0.3. Ethayarajh (2019) showed higher anisotropy for the contextual word embeddings, but that is because it was used BERT-base-cased, which, as we will be able to see, has a higher anisotropy than its uncased counterpart.

Finally, the average pooling is highly anisotropic and has an overall decreasing trend with the layers depth, which confirms the results of Jiang et al. (2022). This higher anisotropy at the first layers can be caused by the stopwords; the contextuality in the first layers is low, which means that the self-similarity (understood as the similarity of the embeddings for the same token in different contexts) is high, and this combined with the high frequency of these tokens, can have a big effect on the average, moving it towards the high frequency hypercone, and increasing the average similarity between sentence embeddings in lower layers. The isotropy in the last layer for the 'none' pooling (token level isotropy) is consistent with the results shown in 3.1, with the average being near 0.2.

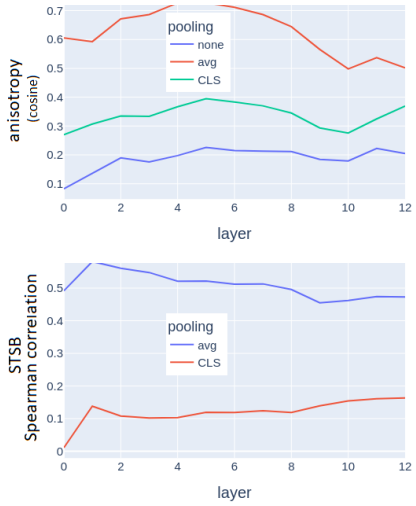

Figure 5: Average cosine similarity (top) and accuracy in STSB (bottom) for different pooling strategies.

3.2.2 Model comparison

For this experiment, we set the average pooling and compared the different models studied in terms of isotropy and semantic isometry. For BERT-base-cased, we inputted uncased text for it to have the same input as the other models. We study the difference of using cased and uncased text in a later experiment. Results are shown in Figure 6.

It is interesting to observe how BERT-base-cased performs clearly better in STSB than BERT-base-uncased, while being around 50% more anisotropic. In addition, we observe that the supervised variant of SIMCSE, the model with the best semantic isometry of the ones analysed, has an anisotropy only slightly below the one of BERT-base-uncased, the less semantic model, and far above the unsupervised variant of SIMCSE. These observations reinforce our hypothesis that, contrary to popular belief, the anisotropy is not the cause of the poor performance of pre-trained Transformer embeddings in semantic tasks.

Another finding is that, although one might expect the embeddings from fine-tuned models to be more semantic than their non fine-tuned counterparts across all the layers, SIMCSE models have very similar semantic isometry to the ones of the base models in the lower layers. For example, in unsupervised SIMCSE, the semantic isometry starts decreasing after the first layer as in the BERT-base models. It is only around the 9th layer when the embeddings make a big shift towards a more semantic space. Indeed it seems that the contrastive learning is mainly acting over the last few layers. This

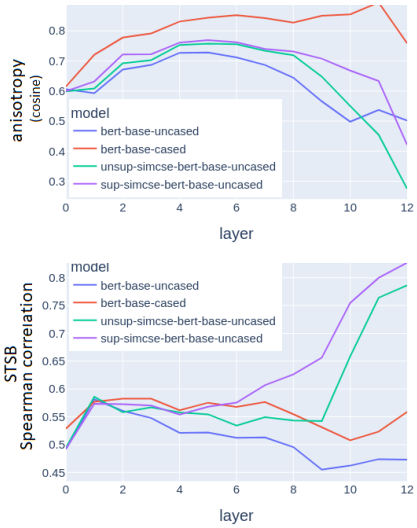

Figure 6: Average cosine similarity (top) and accuracy in STSB (bottom) for different models.

can make sense if we consider that the semantic information is already present in these Transformer language models and that the contrastive learning approach is basically removing all the non semantic information (noise, biases, syntax, etc.), and extracting the semantic information so that it can be reflected through cosine similarity. We hypothesise that the BERT-base language model contains semantic information because it achieves remarkable results in semantic tasks, like in the case of cross-encoders in semantic textual similarity, by just fine-tuning the model with a small dataset.

3.2.3 Bias removal

To continue with our experimentation, we tried removing different sets of output token embeddings related to the biases that we are studying. We apply a similar approach to counterfactual invariance in causal inference (Feder et al., 2021). Specifically, we removed the embeddings from stopwords, subwords, and punctuation marks, with the objective of highlighting the frequency, subword, and punctuation biases reported by Jiang et al. (2022). We also remove CLS, and SEP embeddings, as we have seen that the CLS embedding has poor semantic isometry, and SEP should not contain any relevant information in inferences with a single sentence. It is important to note that we did not remove these tokens from the input sentence. , as that could affect the ability of a language model, trained with syntactically correct sentences, to understand it. Instead, we removed the embeddings after they have been computed by the model, just before the pooling step, like it is performed by Jiang et al. (2022); Yan et al. (2021). That way, we could see how much was each of these token categories contributing to the low isotropy and semantic isometry observed in BERT.

Again, we used average pooling. For the sake of simplicity, we only display the curves for BERT-base-uncased and unsupervised SIMCSE, but the ideas extracted from these experiments also apply to the other models. We show a curve for the removal of each of these categories of tokens individually and for all of them combined, in order to see how the improvements stack and how far we can arrive in terms of semantic performance with this method. The results are shown in Figure 7. It can be observed how the removal of these tokens, individually and combined, improves the results over STSB to different extents in the lower layers of both models. The fact that the effect of removing these tokens is very similar in the first 9 layers of BERT and SIMCSE, reinforces our claim that contrastive learning is mainly modifying the last layers of the network. The only exception to this trend are subwords whose removal is slightly detrimental in both models. This was predictable if we think that subwords are sometimes the result of splitting words with a high semantic load, that are OOV for being very specific, For example, ‘tofu’ is a OOV word and is splitted in ‘#to’, and ‘#fu’. Even if these subwords don’t make sense separately, the attention mechanism of Transformers combines them to get the meaning of the whole word. If we were to remove them from a sentence, it would probably have a high negative effect on its meaning. In fact, the low impact of removing subwords is only due to these tokens being relatively infrequent. Furthermore, our experiments in Section 3.1 showed that the subwords were not biased in the BERT-base model, but they were in SIMCSE. However, this bias does not seem to be affecting SIMCSE, as the removal of subwords also decreases its semantic performance.

In the upper layers of the BERT-base model, the improvement in semantic isometry is still significant, while in the SIMCSE model the curves converge. This second part is surprising, because the SIMCSE model has been fine-tuned for taking into account all the tokens, including stopwords, punctuation marks, subwords, CLS and SEP. We were expecting a higher decrease, especially for the removal of the stopwords, and this can indi-

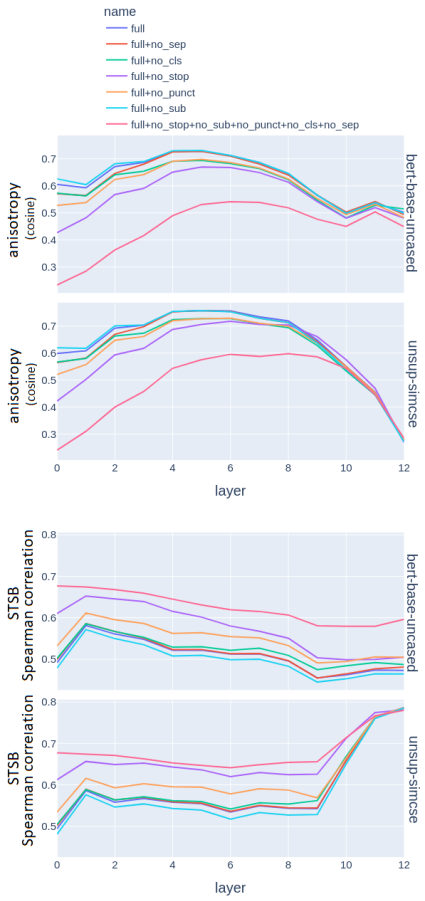

Figure 7: Average cosine similarity (top) and accuracy in STSB (bottom) for the removal of different kind of tokens.

cate that the average contribution of these tokens to the semantics is low, even in contextual word embeddings. However, probably the information given by these tokens had already been distributed throughout the layers via self-attention, so the invariance of the semantic isometry despite their removal from the final average does not necessarily imply that they are not being taken into account for obtaining the sentence meaning. On the other hand, removing the biased tokens decreases anisotropy, but this decrement is significant only in lower layers, especially for stopwords. This confirms the high anisotropy of initial layers in average pooling being partially due to the high frequency of low contextual stopword embeddings. In the higher layers of the BERT-base model, the slight increase in isotropy does not correspond in magnitude to the big increase in semantic performance. Even in the lower layers, where these two metrics improve, this only confirms that biases can generate anisotropy, but not necessarily the other way around. To sum up, we have rejected subword bias and confirmed frequency and punctuation bias. Nonetheless, we still don't know if the punctuation bias is just due to the high frequency of some of the punctuation marks, in which case, the punctuation bias would be contained within the frequency bias. However, we leave this experimentation as future work.

3.2.4 Case removal

With our next experiment we wanted to test how much is the information of the case contributing to the fact that the BERT-base-cased has a way superior semantic isometry than BERT-base-uncased. We tried inputting the text to bert-base-cased in its original form with uppercase words and converted to lowercase. The results are shown in Figure 8.

It can be observed that the uncased embeddings are slightly more anisotropic than the cased ones. This was expected, given the results of the previous section, when we analyzed the case bias. Furthermore, the anisotropy when using a cased input is still much higher than the one of the uncased model. These results match the ones observed in the top plot of Figure 1 (BERT-base-uncased) and in Figure 4 (BERT-base-cased), where the distributions of the cased model show higher values than the ones of the uncased model. This can make sense if we consider that, during the cased model training, most of the lowercase tokens were probably grouped in a hypercone, separated from the uppercase ones, while for the uncased model, this process did not happen because all the tokens were processed as lowercase.

This is another example where more anisotropy does not mean worse semantics. In this case, the more anisotropic variant (the cased model) happens to be more semantic.

What was unexpected to a certain degree was the big drop in semantic isometry when using cased text. This model had been trained with cased text, and the fact that its embeddings are much more semantic with uncased text further proves the idea of biases being a big part of the lack of semantic isometry of contextual word embeddings, and that there could be other unknown biases responsible for this. This big increase in semantic isometry when using the cased model with uncased text is not reflected in any way in the isotropy, which remains roughly the same. This dissonance between both metrics adds evidence in the direction of demonstrating our hypothesis of anisotropy not being the root cause of the lack of semantic isometry.

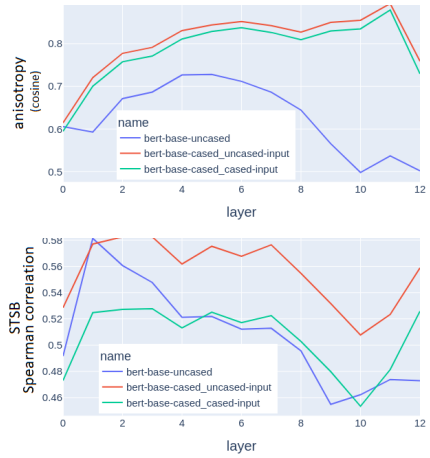

Figure 8: Average cosine similarity (top) and accuracy in STSB (bottom) for bert-base-cased with cased and uncased text.

4 Conclusions

In this paper we have carried out a set of experiments intended to confirm and measure known biases, and to understand their impact over anisotropy and semantic isometry in fine-tuned and non fine-tuned BERT models.

In our results we have not found a clear correlation between isotropy and semantic isometry. In fact, models or pooling methods with a higher anisotropy are sometimes more semantic than others that are more isotropic. However, there exists a correlation between the biases and the semantic isometry. These biases are present in the embedding space, encoding information that is not semantic, like the frequency of a token or its case. This non semantic information distorts the embedding representation space, which leads to poor performance on semantic tasks. Due to this distortion, biases naturally contribute to anisotropy, so there is, in fact, sometimes a certain correlation between isotropy and semantic isometry. But this correlation is spurious and comes from both the anisotropy and poor semantic isometry being a consequence of a high bias. We don't think that there is a causality relation between isotropy and semantic isometry. This means that isotropy correction methods will not achieve substantial improvements over their base models, which has been recently proven by Ding et al. (2022). Therefore, it could be said that assuming that the lack of isotropy of the embedding spaces is the cause of the lack of semantics is a post-hoc fallacy.

Methods that correct the embedding space to a certain degree, like the ones based on contrastive learning, also decrease anisotropy as side effect of removing biases, but we can't expect it to work the other way around, which is, to remove biases by increasing isotropy, as we could just be opening the general cone, while keeping the same islands with the same biases inside. Even after manually removing the known biases (frequency, subword, case, puntuation marks), we are still far away from state-of-the-art unsupervised contrastive learning models. We think this indicates that there are a series of biases that we still don't understand. One promising line of work, taking into account the results from (Luo et al., 2021) and (Kovaleva et al., 2021), could be to try to find a positional bias, and a way to correct it.

Limitations

The main limitation to be mentioned in relation to this work is that we do not produce any improvement over the state of the art in unsupervised nor supervised semantic sentence embeddings. Rather than that, we have focused our research on trying to improve our understanding of BERT embeddings space, and how their isotropy correlates with their semantic isometry. We hope that our results will give some valuable insights to other researchers. Part of our experimentation was already done by Ethayarajh (2019) and Jiang et al. (2022), however, they both used different models and pooling strategies, so their results seem contradictory. Part of our contribution is to match these results and give a more complete picture of the problem, hypothesising that the finding of new biases will contribute to the objective of understanding the lack of semantics in Transformer language models.

Acknowledgements

The authors would like to thank anonymous EMNLP'2022 reviewers for their valuable suggestions to improve the quality of the paper. This work was possible thanks to Voicemod S.L., and partially supported by MCI/AEI/FEDER, UE DOTT-HEALTH (PID2019-106942RB-C32) and FairTransNLP (PID2021-124361OB-C32) projects, and LyrAIcs project through the European Research Council, under the Research and Innovation Program Horizon2020 (Grant agreement No. 964009).

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2020. Isotropy in the Contextual Embedding Space: Clusters and Manifolds.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. 2022. On Isotropy Calibration of Transformer Models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 1–9, Dublin, Ireland. 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.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Models.

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.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. 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 879–895, Online. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. PromptBERT: Improving BERT Sentence Embeddings with Prompts. *arXiv:2201.04337 [cs]*. ArXiv: 2201.04337.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT Busters: Outlier Dimensions that Disrupt Transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional Artefacts Propagate Through Masked Language Model Embeddings. 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 5312–5327, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. *arXiv:2105.11741 [cs]*. ArXiv: 2105.11741.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped Unsupervised Sentence Representation Learning. 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 5168–5180, Online. Association for Computational Linguistics.