UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA (UNED)

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

IS ANISOTROPY REALLY THE CAUSE OF BERT EMBEDDINGS NOT BEING SEMANTIC?

Trabajo de Fin de Master para el Máster Universitario en Investigación en Inteligencia Artificial

Autor:
Alejandro Fuster Baggetto

Director:
Víctor Fresno Fernández

Convocatora de Septiembre, curso 2021-2022

Contents

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

Alejandro Fuster Baggetto Víctor Fresno Fernández

August 2022

Abstract

We conduct a set of experiments aimed to improve our understanding of the lack of semantic isometry (correspondence between the embedding and meaning spaces) of contextual word embeddings of BERT. 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 anisotropy and semantic isometry are a set of biased tokens, that distort the space with non semantic information. For each bias category (frequency, subword, punctuation, and case), we measure its magnitude and the effect of its removal. We show that these biases contribute but not completely explain the anisotropy and lack of semantic isometry of these models. Therefore, we hypothesise that the finding of new biases will contribute to the objective of correcting the representation degradation problem. Finally, we propose a new similarity method aimed to smooth the negative effect of biased tokens in semantic isometry and to increase the explainability of semantic similarity scores. We conduct an in depth experimentation of this method, analysing its strengths and weaknesses and propose future applications for it.

***Keywords*—** semantic textual similarity, sentence embeddings, transformers, natural language processing, deep learning

1 Introduction

For years, natural language processing (NLP) was dominated by recurrent neural networks (RNNs) and its variations like Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks. The capacity of these autoregressive architectures for modelling one dimensional sequences and keeping a relatively long (not as long like it is required in some signal processing applications, for example) context made them especially suitable for NLP.

One of the most common form in which we used to find these kind of models was sequence to sequence. Used a wide variety of applications like machine translation (MT), sequence to sequence models would typically consist on two separate RNNs: The first one, called encoder, would process the entire input and encode it in a single vector, that would be passed to the second one, called decoder, that would predict the output sequence.

Despite their success, these models have a drawback: The decoder usually lacks information for producing the whole output sequence, as all it is receiving is the last output of the encoder.

The solution to this came in form of a mechanism called attention, that would allow the decoder to look at each of the outputs of the encoder and assign then a weight for a weighted sum at each decoding timestep. This way, the decoder would choose its own context at each step considering the full input sequence.

This attention mechanism, that we now call encoder-decoder attention, perfectly complemented sequence to sequence models, but soon researchers realized that RNNs themselves could be replaced by a variation of the attention mechanism called self-attention, and that is how the transformer Vaswani et al. (2017) was introduced.

In each self-attention layer, each timestep has access to the whole input sequence, which removes the necessity of carrying context vectors like RNNs do.

In spite of the substitution of RNNs for self-attention layers, the vanilla transformer is still a sequence to sequence model, with an encoder, a decoder and an encoder-decoder attention.

However, the application of transformers does not end in sequence to sequence tasks like MT. Some transformer models like BERT, take just the encoder part for tasks like classification, while others like GPT, take just the decoder part for generative tasks.

The success of transformer models is partially due by the fact that they scale very well with their number of parameters and the size of the training dataset. The advances in hardware acceleration have enabled the creation of big transformer models, that train over large collections of unsupervised data for general language modeling tasks (LM) like next token prediction or mask prediction. These models have shown impressive results and generalization capabilities.

The Transformer architecture has had an enormous impact over Natural Language Processing to the point that the state of the art on many downstream NLP tasks has been

pushed by finetuning these models with a small supervised dataset.

Semantic text similarity (STS) is a regression task that consists on giving a semantic similarity score between zero and five for a pair of sentences, where a score of zero means that the two sentences have nothing in common semantically speaking, and a score of five means that the two sentences mean the same or are paraphrases of each other.

STS is one of these tasks, where a new state of the art has been set by just adding a linear layer on top of BERT Devlin et al. (2019) (a famous encoder transformer model trained for masked language modelling) and finetuning it to get the similarity score of two sentences concatenated with a separator.

However, this approach (called cross-encoder) has its own problems: Its inference time is too long to be acceptable in a lot of applications. The reason for this is that it performs full self-attention over the input, so it scales quadratically with the length of the concatenation of both sentences. Most importantly, it has to reevaluate everything for each pair of sentences, which stops applications like semantic search or semantic clustering from being feasible.

In order to solve the problem of efficiency of cross-encoders, a lightweight approach (called bi-encoder) consisting on obtaining the distance of two sentences in the embedding space has emerged. In this case, each sentence is passed separately to the transformer, which obtains a semantic embedding for it. Then, obtaining the similarity between sentences is as simple as comparing their embeddings.

Unfortunately, this approach does not work well with embeddings obtained by vanilla pretrained Transformers; Reimers and Gurevych (2019) showed that the average of the output contextual BERT embeddings of the words in a sentence performs poorly as a sentence embedding in semantic tasks. Even the average of the static Glove embeddings Pennington et al. (2014) of the words in a sentence, often used as a weak baseline, results in a more semantic sentence embedding, despite their lack of contextuality.

On the other hand, studying this mismatch between the word embedding and meaning spaces, Gao et al. (2019) diagnosed high anisotropy in Transformer Language Models. This means that the embeddings do not follow a uniform distribution with respect to direction, or, what is the same, they concentrate in an hypercone instead of ocupying the whole space. This behaviour is anomalous, as one expect transformer to try to make use of all its representation power.

This anisotropy has been named as the representation degradation problem and could be closely related with semantics to the point that some authors have tried to correct the semantic isometry, understood as the correspondence between the meaning and embedding spaces, of transformers by applying isotropy correction techniques Li et al. (2020); Su et al. (2021). Other approaches, based on contrastive learning have also recently achieved remarkable improvements in semantic isometry Zhang et al. (2021); Giorgi et al. (2021); Yan et al. (2021); Gao et al. (2021).

Contrastive learning methods have been naturally successful in correcting embedding

spaces because they pull together semantically similar sentences and push apart semantically dissimilar ones.

Despite their good results, these techniques have also their own problems. The main difference between these methods is how they perform the selection of the positive (semantically similar) and negative (semantically dissimilar) pairs necessary for training, and there is where the main difficulty is. For example, Zhang et al. (2021) used back translation from English to German to obtain augmented views of a sentence, or 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 top of previous work.

These contrastive bi-encoder models achieve unprecedented results in terms of semantics, but we don't fully understand the root cause of the lack of semantic isometry observed in pretrained Transformer Language Models. We think that any finding in this area can be very relevant and open new paths of research that can lead to future improvements in bi-encoder methods.

Our main aim in this paper 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 call bias to any information from a sentence that is encoded in the embedding space and that is not relevant to its meaning, such as the length of a sentence.

This paper is structured as follows. In Section 2 we cover the related work. Next, in Section 3 we describe the experiments conducted and present their corresponding result analysis. Next, in Section 4 we draw our conclusions, derived from our empirical results and in section 5 we leave some ideas for future work. Finally, in Section 6 state the limitations of this work.

2 Related work

The anisotropy found by Gao et al. (2019) in Transformer-based Language Models has been called the representation degradation problem. This problem is produced by the combination of the Zipfian nature of natural language, which means that the frequency of the tokens follows a zeta distribution, and the log-likelihood loss function used to train . This degradation makes the most frequent tokens to concentrate in a cone in the embedding space, having a more sparse space for infrequent ones, which decreases isotropy.

Contrastive learning methods seem to correct anisotropy to some extent while greatly improving the semantic isometry. This could make us think that the anisotropy was, in fact, part of the semantics problem.

However, Jiang et al. (2022) realized, through a series of experiments, that there exist certain biases in the BERT model, and that anisotropy is not always equivalent to poor semantic isometry. Although the insights by Jiang et al. (2022) are certainly interesting, they seem to contradict at some degree previous works like Gao et al. (2019); Ethayarajh (2019); Li et al. (2020), that attribute the lack of semantics of transformer embeddings to anisotropy.

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. Indeed, there are few dimensions in BERT embeddings whose module is disproportionally high in comparison with the other dimensions. These outlier dimensions distort the cosine distance, reducing isotropy. Whether or not these dimensions affect semantic isometry is still not verified, but their origin seems to be found in positional encoding, which could mean that there exist a positional bias in these models.

By clustering the transformer token embeddings and standardizing each cluster, Cai et al. (2020) showed that, despite the model 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), who show that isotropy correction techniques don't improve results in most semantic tasks.

If, like Cai et al. (2020) claims, the anisotropy comes from the existence of different clusters, and these clusters encode non-semantic information like token frequency, these clusters could be matched with the biases described by Jiang et al. (2022) and the representation degeneration by Gao et al. (2019), that describes the frequency bias.

As we can see, in the literature there is no consensus in which is 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 chaos 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) does it at the sentence level, by averaging the word embeddings. We therefore find that there is room for research on these aspects.

3 Experimentation and results

Our experimentation proceeds as follows: First, we make an exploratory analysis to better understand the magnitude of the different biases studied. Next, we compare several configurations (pooling strategies, models, bias removal, etc.) in terms of isotropy and

semantic isometry. Finally, we propose a new sentence similarity metric, that we call pairwise similarity, and that works at the token level, which allows us to better understand the nature of BERT's embedding space.

In our experiments, we try the following BERT variants: BERT-base-uncased, BERT-base-cased, unsupervised-SIMCSE-base, and supervised-SIMCSE-base. We include a cased model in order to compare it with its uncased counterpart and 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), as it is a popular contrastive learning bi-encoder model that achieves state-of-the-art results in semantic tasks. 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 performance of the non fintuned bi-encoders.

Finally, note that all the models analysed share the BERT base architecture. Although this choice might seem arbitrary, we base it in the fact that different studies like Gao et al. (2019); Ethayarajh (2019); Kovaleva et al. (2021) have shown that different Transformer-based Language Models have similar behaviours regarding low isotropy and poor semantic isometry, even when they differ in number of parameters, architecture or learning objective.

3.1 Exploratory analysis of biases

The length of a sentence or the frequency of its tokens are examples of well-known biases, as they are non-semantic information encoded in the embedding space. Four kinds of biases are defined by Jiang et al. (2022): frequency, case, subword and punctuation. We understand as subwords, the pieces of words generated by the BERT tokenizer when it encounters out of vocabulary words. For example, the word "embedding" could be splitted in the subwords "#em", "#bed", "#ding". The nature of out of vocabulary words is varied. Sometimes they are named entities, while other times they are uncommon words or variations of other words like the plural of a noun or the past form of a verb.

All of the aforementioned biases could partially overlap with token frequency. For example, lowercase tokens are much more frequent than uppercase ones, some punctuation marks like "," and ".", are more common than normal words, and some subwords like "#s" (from plural) are extremely 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 cone due to the combination of the Zipf distribution of frequency in natural language and the log-likelihood loss function used in training.

For having a perception of how severe are each of these biases in semantic and non-semantic embedding spaces, we sample 1000 random sentences from the Wikipedia corpus and we plot the distributions of the similarities between pairs of embeddings of different kind of tokens. For all the plots we use cosine as the similarity metric and the last layer word embeddings from BERT-base-uncased and unsupervised-SIMCSE-base, except for

the case bias, where we only use BERT-base-cased. We expect to find lower biases for SIMCSE, as its semantic isometry has been shown to be higher than the non finetuned BERT.

3.1.1 Frequency bias

For the Frequency bias, we first decided to compare the most frequent tokens with less frequent ones. Howver, we finally opted to use a list of stopwords instead. The reason for this is that we saw that the list of the most frequent tokens is mainly formed by stopwords, punctuation symbols and some very common nouns like "man" or "woman". We treat punctuation symbols as a separate category because, despite some of them being very frequent, we do not want to preassume that there is not a 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 context and whose contribution is predominantly syntactic. In section 3.2 we remove the biased tokens to see the effect in isotropy and semantic performance and we think that removing nouns, even if they are frequent and biased, can 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.

In Figure 1 we see that, even in the last layer of BERT, where stopwords are supposed to be very contextual, as proved by Ethayarajh (2019), the average similarity between these words is still slightly higher than the similarity between stopwords and less frequent words, or between less frequent words. This confirms the frequency bias and relates it with the fact that frequent tokens are concentrated in a cone in the vector space, while less frequent tokens are more sparse. We can also see that this gap is reduced for the SIMCSE model, that seems to have corrected most of this bias trough contrastive learning.

3.1.2 Subword bias

The contrary happens in the case of subword bias. In the Figure 2 we can see that the subword bias, understood as the gap of the green distribution (similarity between pairs of stopwords) and the other two, is significantly higher in the SIMCSE model, despite its overall average cosine similarity being smaller (or more isotropic) than in the base model.

3.1.3 Punctuation bias

The case of punctuation marks is a little bit trickier. For the base model, punctuation marks tend to be more sparse in average than whole words, as we can deduce in Figure 3 by the high variance of the green distribution in the base model, that represents the distance between punctuation marks. Furthermore, we can see a cluster at the right that reaches very high similarities, being some of them near one. Although these high

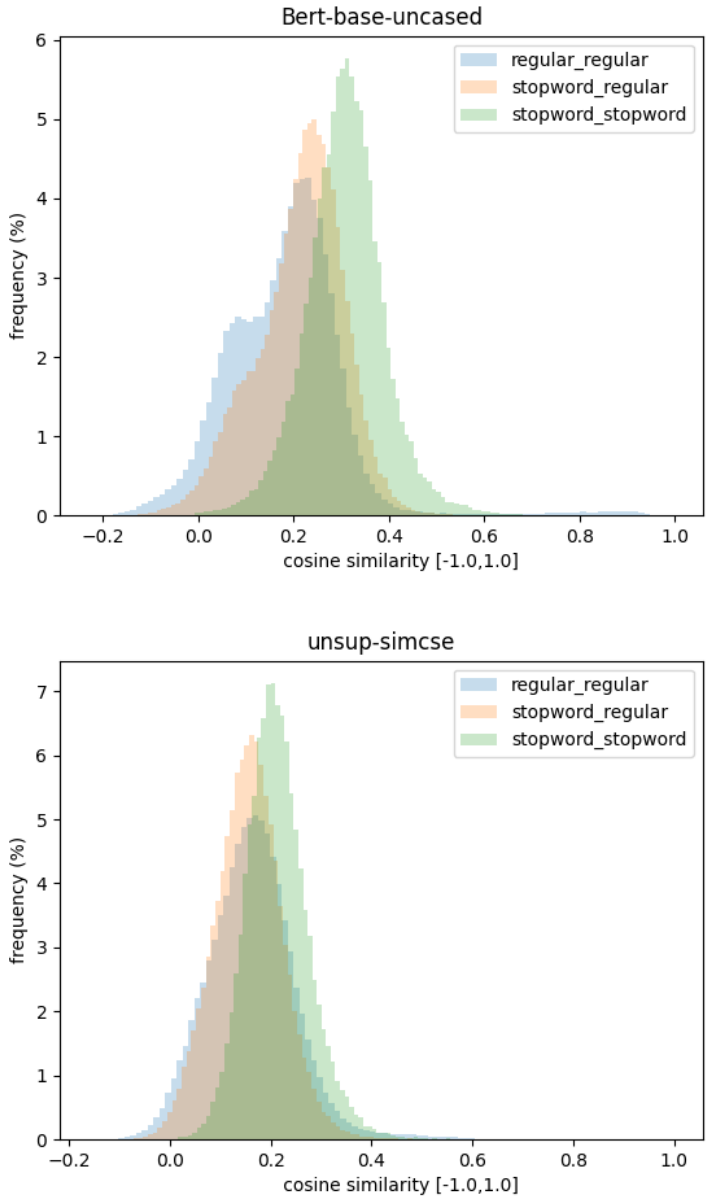

Figure 1: Average cosine similarity between stopwords and other tokens from uncased BERT (top) and unsupervised SIMCSE (bottom)

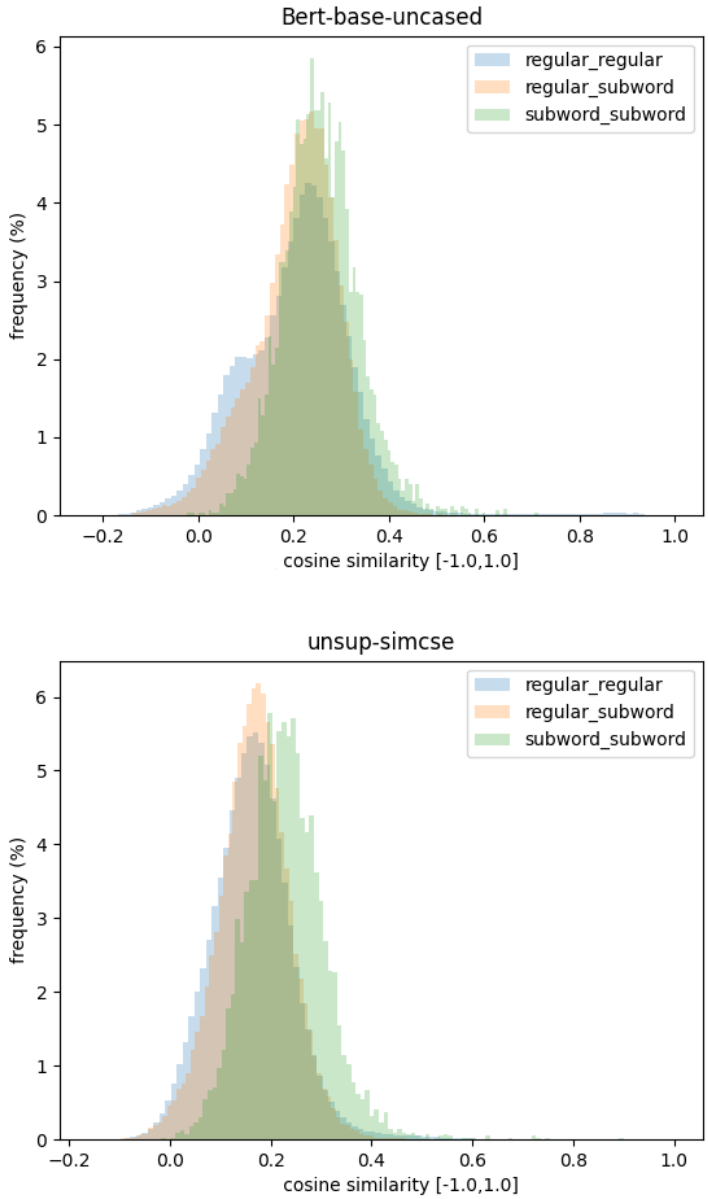

Figure 2: Cosine similarity between words and subwords from uncased BERT (top) and unsupervised SIMCSE (bottom).

similarities between contextual embeddings are somewhat surprising, it all could come down to frequency. Some punctuation marks, like comas or dots, are extremely frequent, while others like exclamation or question marks are also relatively frequent, and others are very infrequent, like asterisks or slashes. All of this, combined with the representation degradation problem (or frequency bias) that we have 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 managed to solve 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 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. One is the named entities, and the other is tokens that lay at the beginning of a sentence. Obviously these two groups are not mutually exclusive, as we can have a named entity in the beginning of a sentence. Both types of tokens have varied frequencies. For example, there are very common named entities like months, days, or names of countries or celebrities. There are also words that are very common as sentence beginners like "The". We think that the small peaks in the right of the distribution could be due to these 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 Experiment conclusions

Take into account that most of these points are just hypothesis that would explain the results, but that require verification. What we can conclude with certainty, though, is that, in the non finetuned BERT models there is indeed such thing 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 are situated in a cone 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, we can see that, in general, there is a better overlap of the different distributions. Furthermore, the distributions tend to have a lower variance and and be centered close to zero. The mean value of the distribution being around zero tells us that these models are more isotropic. However, we think that this is not the important point. Even if the distribution was centered

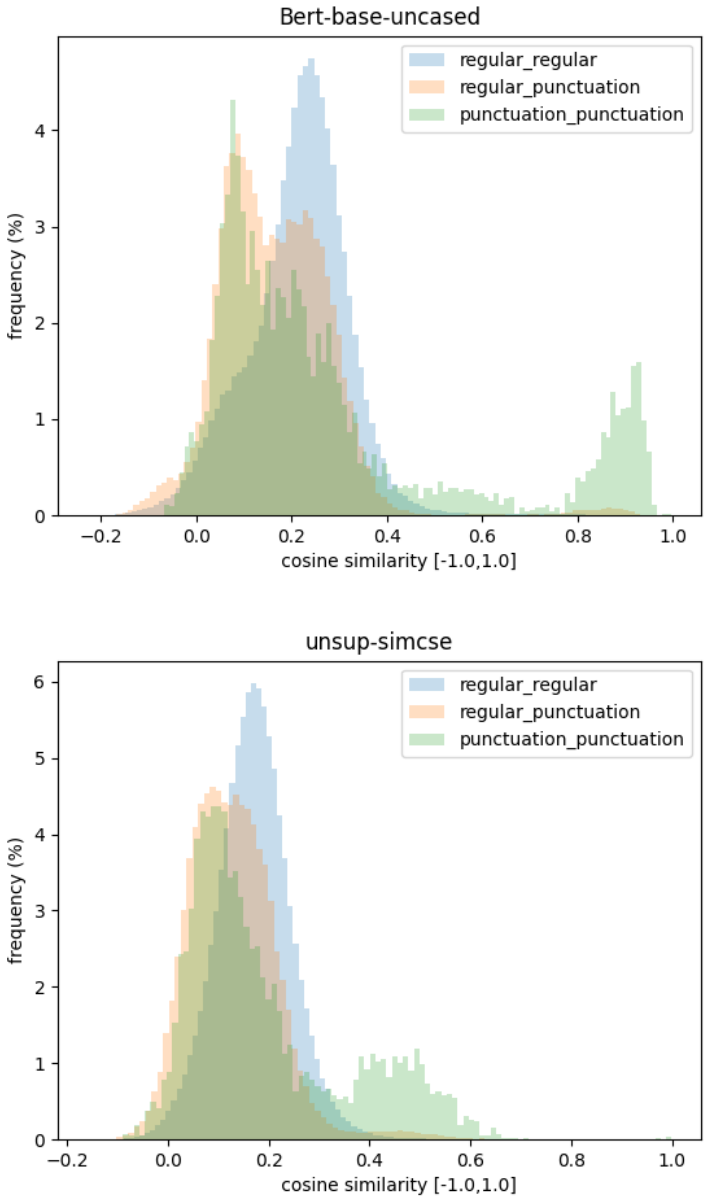

Figure 3: Cosine similarity between punctuation marks and other tokens from uncased BERT (top) and unsupervised SIMCSE (bottom)

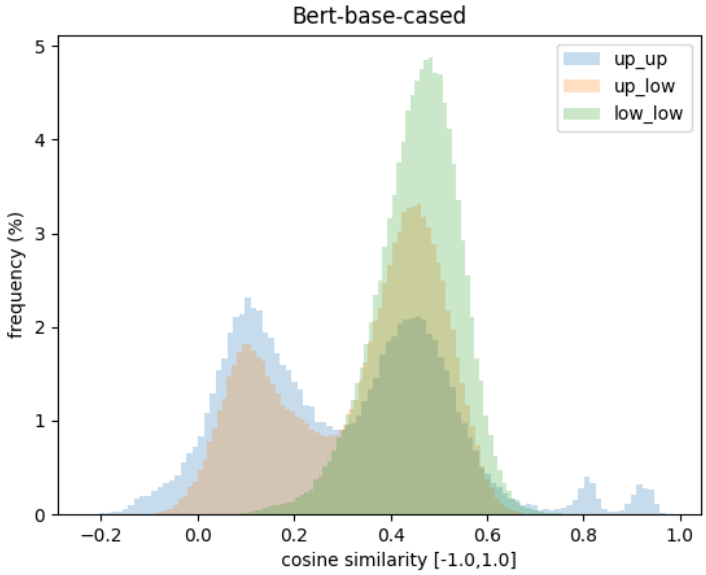

Figure 4: Cosine similarity between uppercase and lowercase words from cased BERT.

in 0.8, that would only increase anisotropy and would mean that all the embeddings lay in a narrow cone, 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 was semantic and not biased, being concentrated in a narrow cone would, by definition, decrease the isotropy and potentially the representation power of the model, but should not distort the semantic isometry. If a pair of tokens is semantically closer than other, this relation would be kept even if we reduce the angle of the representation cone.

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 proved 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; as already mentioned, we understand semantic isometry as the correspondence between the spaces of embedding and meaning.

For isotropy evaluation, we sampled 10,000 random sentences from the Wikipedia corpus and compute the cosine similarity between them in the embedding space. For semantic isometry, we used the test set from the semantic textual similarity benchmark (STSB) Cer et al. (2017) and compute the Spearman correlation between the cosine similarity obtained for each pair of sentence embeddings and their annotation. We conduct the following experimentation for all the layers of each model, as the literature tell us that different layers store different kinds of information.

3.2.1 Pooling comparison

First, we compare different pooling strategies for the bert-base-uncased model. We use 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 token average in semantic tasks for the non finetuned BERT models. However, we still think that it is worth to include this strategy in our experimentation, especially to see how it behaves in terms of isotropy.

Additionally, we include "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, we are computing the cosine similarity between pairs of token embeddings, instead of sentence embeddings. This can only be done for isotropy, as the benchmark for semantic isometry (STSB) is only available between pairs of sentences. We can see the results of this experiment in Figure 5.

First of all, as it was previously stated, we can observe that the semantic isometry of CLS pooling is much worse than the one of average pooling, despite being much more isotropic than the average pooling.

On the other hand, tokens (none pooling) are not that 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 they used bert-base-cased, which, as we will 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, 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 cone, and increasing the average similarity between sentence embeddings in lower layers.

The isotropy results in the last layer for the "none" pooling (or token level isotropy) are consisten with the results showin in 3.1, with the average being near 0.2.

3.2.2 Model comparison

For this experiment, we fix the pooling method to the average and we compare the different models studied in terms of isotropy and semantic isometry. For Bert-base-cased, we input 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 see how BERT-base-cased performs clearly better in STSB than bert-cased, 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 pretrained Transformer embeddings in semantic tasks.

One final trend that can be observed here is that, although one might expect the embeddings from finetuned models to be more semantic that their non finetuned counterparts across all the layers, SIMCSE models have very similar isotropy semantic isometry to the ones of the base models in the lower layers. Taking unsupervised SIMCSE as an example, its semantic isometry starts decreasing after the first layer as in the base models. It is only around the 9th layer when its embeddings make a big shift towards a more semantic space, as it is reflected in the plot. Indeed it seems that the contrastive learning is mainly acting over the last few layers.

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

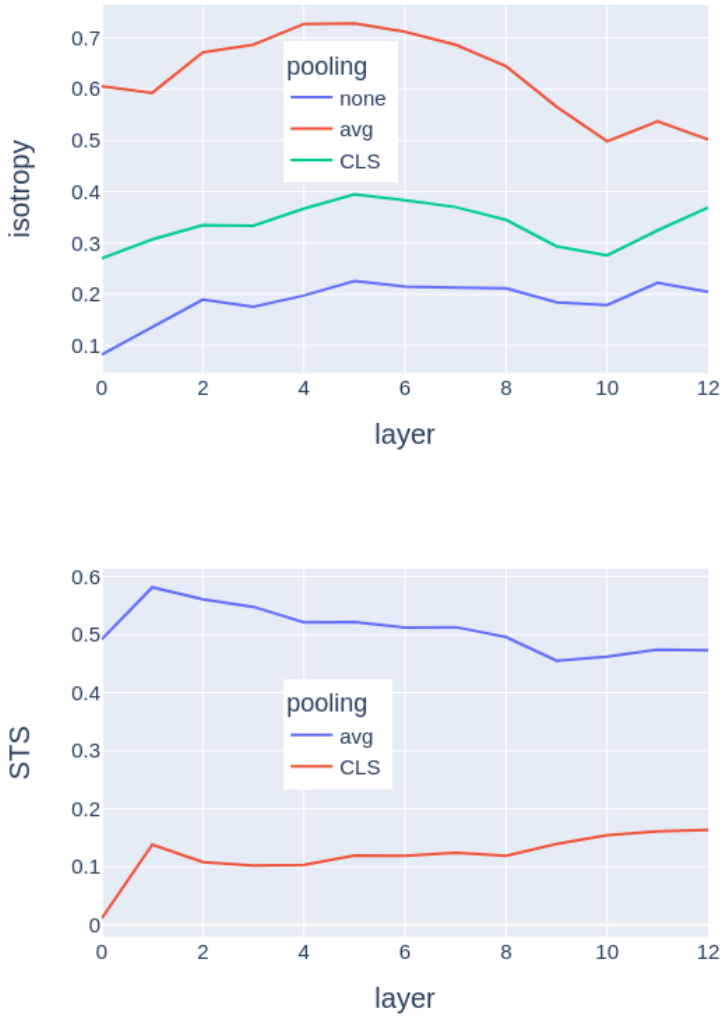

Figure 5: Average cosine similarity (top) and accuracy in STSB (bottom) for different pooling strategies.

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

3.2.3 Bias removal

To continue with our experimentation, we try removing different sets of output token embeddings related to the biases that we are studying to see how they affect to the isotropy and semantic isometry of the model. Therefore, we apply a similar approach to counterfactual invariance in causal inference Feder et al. (2021). Specifically, we remove 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, 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 do 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 remove the embedding after it has 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 can see how much are each of these token categories contributing to the low isotropy and semantic isometry observed in BERT, and whether or not they are causes to the lack of them.

Again, for these experiments we use average pooling. For the sake of simplicity in the figure, 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 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.

For the most part, we see that the removal of these tokens, individually and combined, improves the results over the semantic similarity benchmark (STSB) to different extents in the lower layers of both bert-base-uncased and unsupervised-SIMCSE.

The fact that the effect of removing these tokens is very similar in the first 9 layers of both models, reinforces our claim that contrastive learning is mainly modifying the last part of the network.

The only set of tokens whose removal does not improve the semantic isometry is subwords. We can see that removing subwords is slightly detrimental in both models. This was predictable if we think that subwords are sometimes the result of words with a high semantic load, that are outside of the vocabulary for being very specific. Following the example given above, the word "tofu" is out of the vocabulary and is splitted in "#to", and "#fu". Even if these two pieces don't make sense separately, the attention mechanism of transformers combines both to get the meaning of the whole word. If we were to remove these two subwords from a sentence, it would probably have a high negative effect on its meaning. In fact, the low impact of removing subwords that we see in the figure, is only due to these tokens being relatively infrequent. Furthermore, our experiments in Section 3.1 showed that the subwords are not biased in the base model, but they are in

SIMCSE. However, this bias does not seem to be affecting SIMCSE, as the removal of subwords also decreases semantic performance in this model.

In the upper layers of the 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 finetuned for taking into account all the tokens, including stopwords, punctuation marks, subwords, CLS and SEP. However, the removal of these tokens only decreases the Spearmann correlation with the STSB gold standards in 0.005. We were expecting a higher drop, especially for the removal of the stopwords. This can indicate that the average contribution of these tokens to the semantics of a sentence is, in general, low, even in contextual word embeddings. However, probably the information given by these tokens has 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 necesarily 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 we can see this decrement is significant only in lower layers, especially for stopwords. This confirms our statement before about 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 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. If this was the case, there would be no such think as a punctuation bias, as it would just be contained in the frequency bias. We know that removing frequent punctuation marks would improve the semantic isometry, but we want to know if all the improvement of removing punctuation marks comes from there, or if removing less frequent punctuation marks also contributes to the overall score.

For this, we have elaborated the following list of frequent punctuation marks: [".",",","’","-",":"], and removed their output embeddings before the average, as we have done with stopwords and punctuation marks. The results were the same between the removal of all punctuation marks to the removal of the most frequent ones. However, this experiment is not conclusive because the infrequent punctuation marks will have less of an impact on overall scores due to their low frequency in the dataset that we are using. Instead, we should make a test set by choosing sentences that contain plenty of these infrequent punctuation marks, so their effect becomes noticeable. Due to time constraints, we leave this experimentation as future work.

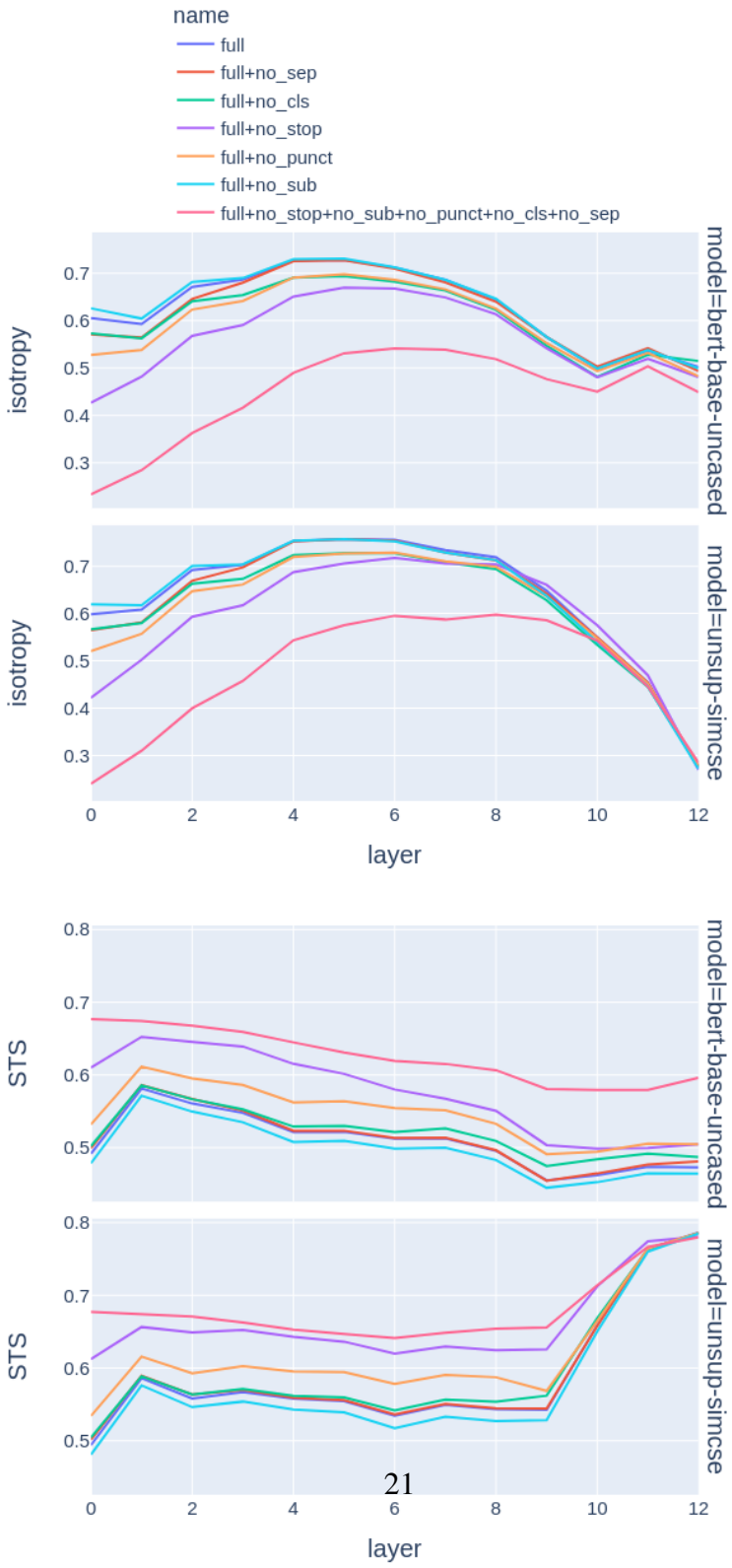

Figure 7: Average cosine similarity (top) and accuracy in STSB (bottom) for the removal of different kind of tokens.

3.2.4 Case removal

With our next experiment we want 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 try 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.

We can see that the uncased word 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. We can confirm this by looking at the top plot of Figure 1, that was made with bert-base-uncased and Figure 4, that was made with bert-base-cased. The distributions present in the second figure have higher values than the ones from the first figure.

This can make sense if we consider that, during the cased model training, most of the lowercase tokens were probably grouped in a cone, 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 happens to be more semantic.

What was unexpected to a certain degree is the big drop in semantic isometry when using cased text. This model has been trained with cased text; 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.

3.3 Pairwise similarity

In previous experiments we have seen how, in the base BERT models, none of both simple pooling strategies give satisfactory results for semantic similarity. In the case of the average pooling we hypothesize that part of the problem is that averaging a sequence of token embeddings that contain biased tokens could be distorting the sentence embedding. For confirming that this is the case, we design the following alternative strategy for computing the similarity that was inspired by the attention mechanism, the basic building block for transformers.

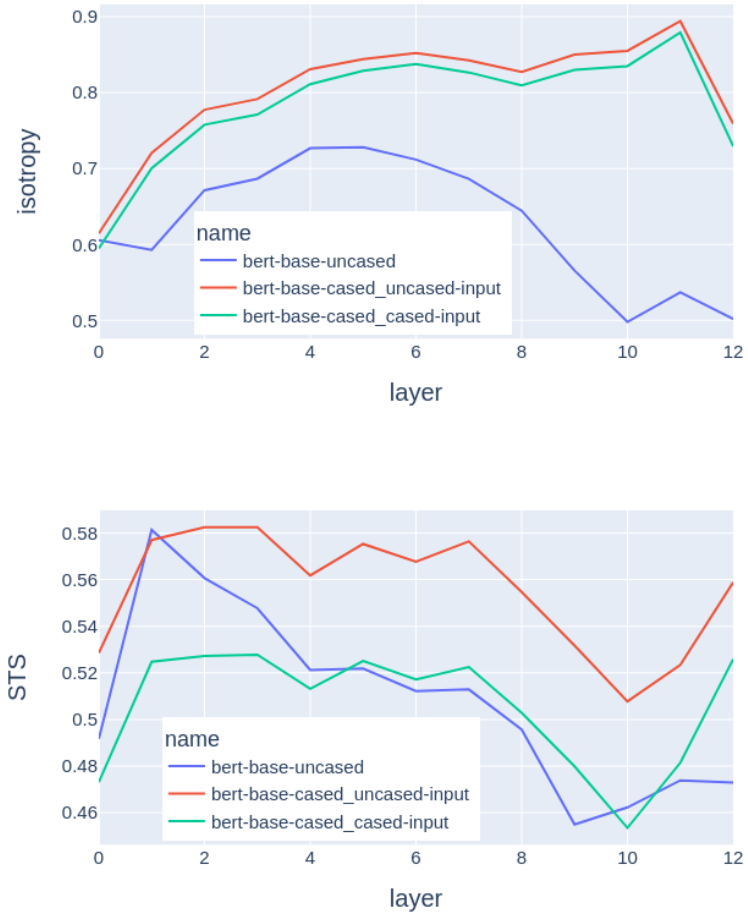

Figure 8: Average cosine similarity (top) and accuracy in STSB (bottom) for bert-base-cased with cased and uncased text.

3.3.1 The method

Given a pair of sentences, instead of trying to directly obtain the sentence embeddings and computing the similarity between them, we keep the contextual word embeddings for both sentences and compute the cosine similarity between each pair of word embeddings. We arrive to a matrix like the one shown in Figure 10, whose values are in the range [-1,1]. We then obtain the maximum values for each row and for each column. These values try to represent the alignments between the tokens of both sentences. Finally, we multiply each of these "alignment" similarity scores by the modules of the two token embeddings involved in it, and compute the average of these values to obtain the similarity score of both sentences.

The reason of computing the alignments from the cosine similarities is that we want to know which tokens are semantically present in both sentences, independently of their semantic weight.

However, as we described above, to obtain the final score, we first multiply the alignments by the embeddings' modules because they represent the amount of information of the tokens, while the angles represent their meanings. If we discarded the module information, we would be giving the same weight to a match in the word "is", than to a match in the word "tofu", while their contribution to the overall meaning of the sentence is very different, being, of course, the second one much more relevant. To sum up, if two sentences contain the word "tofu" they are probably more similar between them than if they both contain the word "is".

Recapitulating the explanation at the beginning of this section, the idea behind this method is that, by not averaging the word embeddings in a sentence, we are avoiding part of the negative effect of biases on semantic performance.

3.3.2 Comparison with bias removal

To prove this hypothesis, we compared the semantic isometry obtained by this method to the one achieved in Section 3.2 by removing the biased tokens. These results are shown in Figure 9, and they are very interesting.

Our hypothesis seems to validate to some degree in lower layers. There, the pairwise similarity obtains similar results (although slightly worse) than the removal of all biased tokens in both models. In addition, we see that, in lower layers, the combination of both techniques only obtains a slight improvement, much lower than the addition of the individual improvements of both techniques. This indicates that there is an overlap between these two methods, and that, with the pairwise similarity, we are at least smoothing the influence of the biases over the semantic isometry of the model.

All of this is true for the lower layers, but completely changes in the last layers. Here we are talking about the base model, as we already know that, due to the way it was trained, the best performance for SIMCSE in the last layers is obtained without doing any

changes. In the last layers of bert-base-uncased, the pairwise similarity, not only does not improve, but is actually slightly detrimental to the semantic isometry. In this case, the combination of both techniques basically overlaps with the removal of biased tokens. For some reason, the high context in more advanced layers is invalidating our hypothesis. In the next section, we will analyse some examples to try to understand this behaviour.

3.3.3 Analysis of alignment matrices

We think that the main value of pairwise similarity is not the increase in semantic isometry of the lower layers shown before, but the fact that it enables a deeper study of each pair of sentences, which helps us understand its outcome. When pooling with the average and doing the cosine similarity between sentence embeddings, a lot of information is lost in the process and it is difficult to know where an unexpected similarity score is coming from. As we are going to see, in comparison, pairwise similarity is much more explainable.

We applied this method to a random pair of sentences extracted from the STSB test set. These two sentences happen to be the following: "One woman is cutting a block of tofu into small cubes", "One woman is slicing some tofu". The annotated similarity score for these pair of sentences is a 4 in a scale from 1-5, which, according to STSB Cer et al. (2017) means "The two sentences are mostly equivalent, but some unimportant details differ". The only semantic difference between these two sentences is that the first one contains the information about the shape of the resulting tofu (small cubes), while the second one does not give this information.

In Figure 10 and Figure 11 we show the alignment matrix of this method for the models bert-base-uncased and unsupervised SIMCSE respectively. The values of the alignment matrices are obtained via cosine similarity, so they are in the range [-1,1]. However, due to the general lack of isotropy of these models that we have discussed earlier, we do not obtain scores lower to -0.2, so we set as as a lower bound for the color map of the figures in order to increase visibility.

In addition, for both models we obtain the alignment matrix for the first and last layers (0 and 12), in order to analyse the difference between static and contextual embeddings.

The first thing we notice in these figures is that, in both cases, the similarity values are generally higher in the contextual embeddings than in the static ones, especially for tokens that should be semantically unrelated. This is even more substantial in the case of SIMCSE, were pairs of tokens like "one" and "is", have a similarity of 0.76 while "slicing" and "cutting", that are synonyms and that are even being used in the same context, have a score of 0.72. One could try to blame the representation degradation problem for this, but we are talking about SIMCSE, a model that has a good semantic performance, a low frequency bias and a very low anisotropy in its last layer, has we have seen in previous experimentation. We started this experiment with a preconceived idea of how the contextual embeddings are, and these results change it completely. We thought that contextual

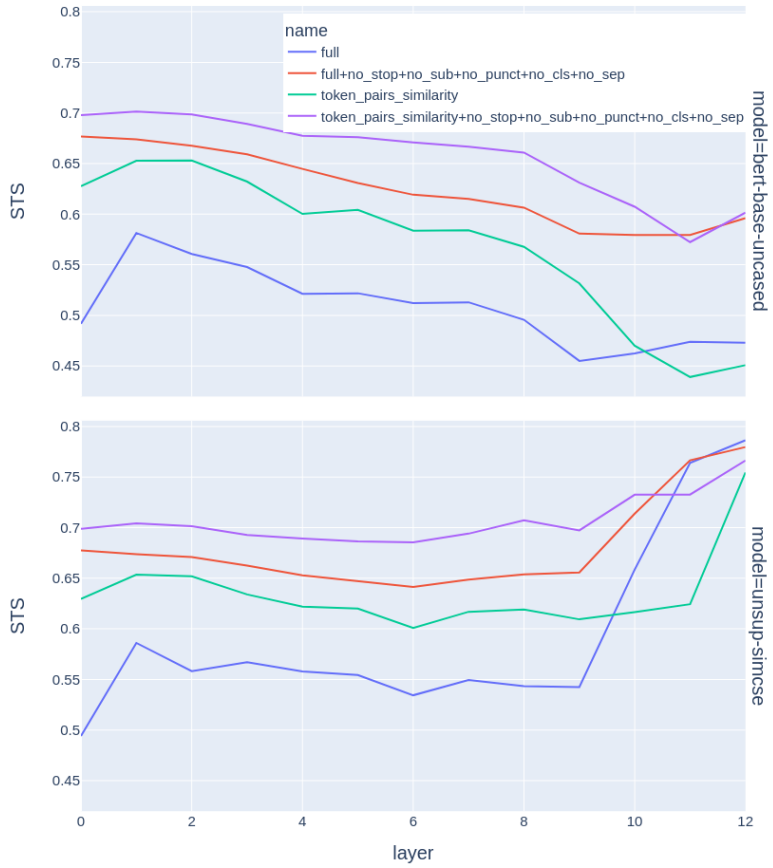

Figure 9: Accuracy in STSB for bert-base-uncased and unsupervised SIMCSE with bias removal, pairwise similarity and both

embeddings would be similar to word embeddings but with the disambiguation of polysemous words, and perhaps the incorporation of some relevant stopwords' information into the words they reference, like "into" in "cutting". However, we didn't think that the context would be a subcone in the space and that all the embeddings in the same context would be located there, despite their individual meaning. In fact, once created the context, the individual words are not that important, as they are all inside that cone. We can see that, even if we remove the word "slicing", that is crucial for the meaning of the first sentence, this information has already been used in the creation of the context and there are a lot of other tokens inside this context, including the CLS, that have high similarities with tokens from the other sentence, so the overall similarity score would not decrease significantly.

Now there is no question of why this pairwise similarity gave poor semantic results in the upper layers. In these layers, the influence of the context is so strong that it doesn't make sense to look at the similarities between individual tokens.

To end our observations about the results of the last layer, we see that, comparing both models, there is a noticeable difference in the similarities of the CLS and "." tokens. The base model fails to include them in the context. It is interesting to see how they achieve very high scores with the CLS and "." tokens of the other sentence respectively, but they achieve low similarity scores with the rest of the tokens. This can be explained by the biases that are more present in the base model than in the SIMCSE one. In Section 3.1 we saw that there are clusters in the space that contain punctuation marks. The embeddings of "." here seem to be sharing one of them. About the CLS, in Section 3.2 we saw that it is anisotropic, which means that its self-similarity is high. We did not measure the isotropy of CLS in SIMCSE models, but, for what we have seen with the average pooling, the anisotropy of CLS is likely low in the last layers. The SIMCSE model not only places CLS in insde the context, but in the center of it, being one of the tokens with higher similarity with all the other ones.

About the results in the first layer, as expected, they are extremely similar for both models. As we have pointed our before, in lower layers there is no context, so self-similarity is high or words, which, in this case, results in very high scores for the words that are repeated in both sentences. On the other hand, we see a score of 0.46 between "slicing" and "cutting", which is not bad, but that gets distorted by scores like the 0.43 between "of" and ".", that are spurious and due to the representation degradation problem.

Given these results, our next idea was to see what would happen with these matrices if we passed a pair of sentences with a very similar syntax and words. Theoretically, we should be able to fool the static embeddings, that do not distinguish between the multiple meanings of words, while SIMCSE contextual embeddings should be able to pass the test by placing both context far away from each other and giving a low similarity score. Therefore, we analyse the following pair of sentences, designed as an adversarial test: "One woman is cutting a block of tofu into small cubes", "One woman is buying

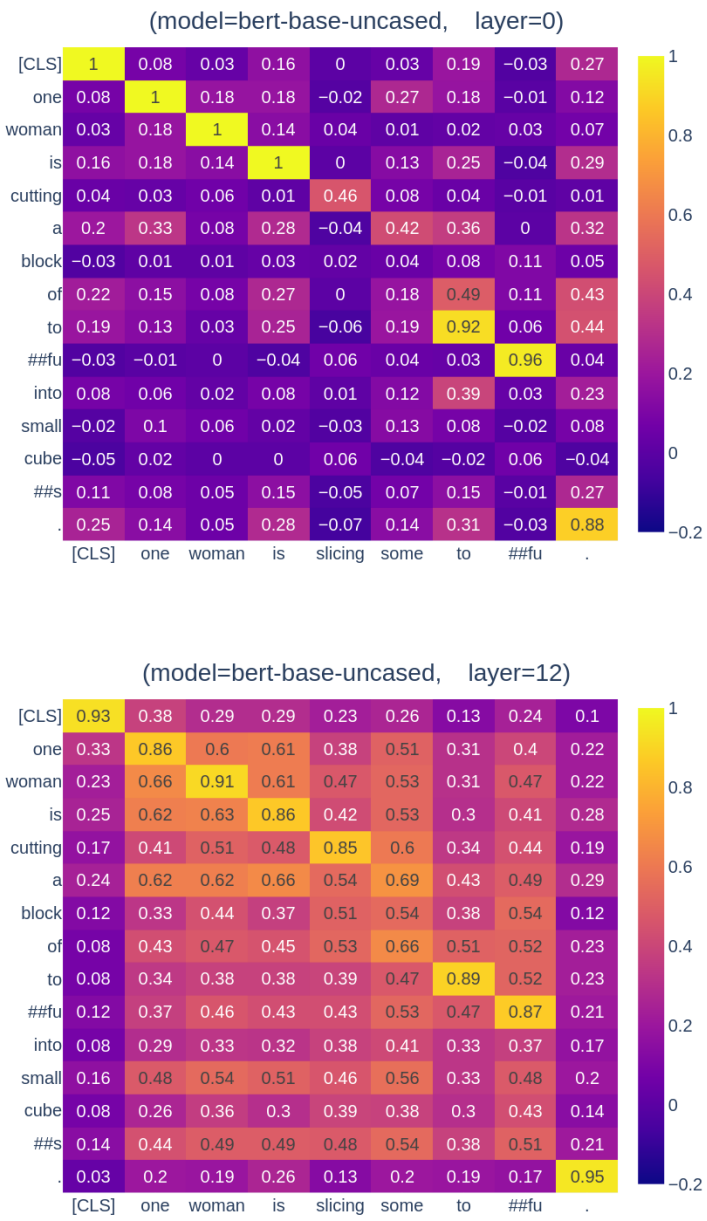

Figure 10: Cosine similarity between the token embeddings of the adversarial sentence pair from the first (top) and last (bottom) layers of bert-base-uncased model

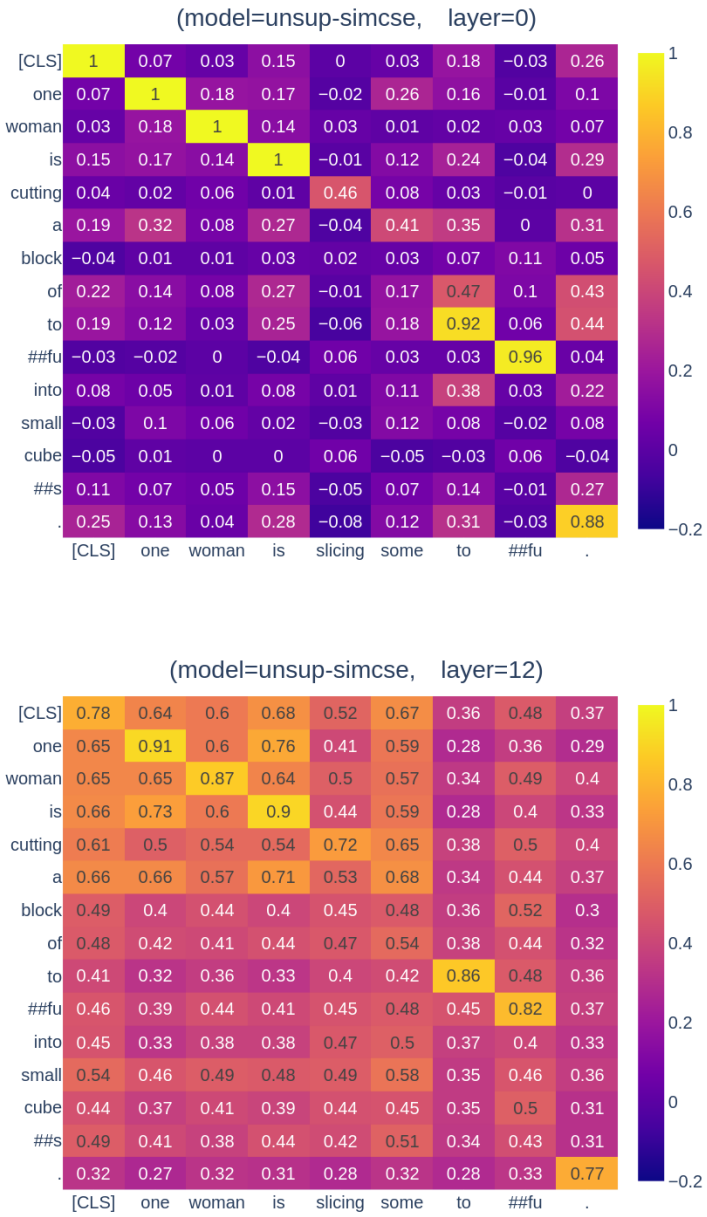

Figure 11: Cosine similarity between the token embeddings of the original sentence pair from the first (top) and last (bottom) layers of the SIMCSE model

a block of apartments", both of which follow the structure "One woman is X a block of Y". The alignment matrices corresponding to this experiment are located in Figure 12 and Figure 13.

For the first layer, as expected, we clearly fool the models into high similarity alignments by using the same words.

For the last layer, it is not that simple. We believe that we have managed to fool both models to a certain degree, as both still assign relatively high scores to alignment pairs of tokens, even when they are not related in context. Of ocurse, SIMCSE, the more semantic model, has lower scores for this sentence pair than the base model, so it has been able to better differentiate between meanings here.

3.3.4 Analysis of pairwise similarities

Despite some interesting information being extracted from the alignment matrix, we also want to compute the final similarity score for each of this cases. This score is not only influenced the strength of the alignments but also the modules of the tokens forming them. We have not conducted any experimentation to explore the modules in different tokens, layers, or models, and we leave that as a future work. For now, lets just assume that the module is related with the amount of information of the word, as it has been shown by the literature.

We obtained the similarity scores for each case, but they don't make sense in an absolute scale. We can get a number like 150, but we don't know if that is actually a lot or not. The scores make sense relative to each other. For measuring to what extent did our adversarial example fool each of the models, we tried a third sentence pair, whose semantic similarity score should be similar to the adversarial one. If the adversarial pair obtains a higher score, we will know that we were able to fool the models by repeating words and syntactic patterns. The pair of sentences is the following: "One woman is cutting a block of tofu into small cubes", "One woman buys a house". We call this pair "non related" as it does not have in common nor the syntax nor the meaning. The pairwise similarity scores for each and the other pairs are shown in Table 1.

In this case we can quantify some of the observations made earlier.

The results of the first layer extremely similar between both models, with SIMCSE always rating a little bit lower. The first layers get fooled by the adversarial example, that achieves a similarity score that is similar to the original one, far above the non related example.

In the last layer, SIMCSE is more accurate than the base model, as it has a higher score for the original pair, and a lower one for the adversarial and non related pairs than bert-base-uncased. In addition, reading this numbers, we can confidently say that the last layers of both models have been partially fooled by the adversarial example. Even SIMCSE, that is a state of the art bi-encoder model for unsupervised semantic textual

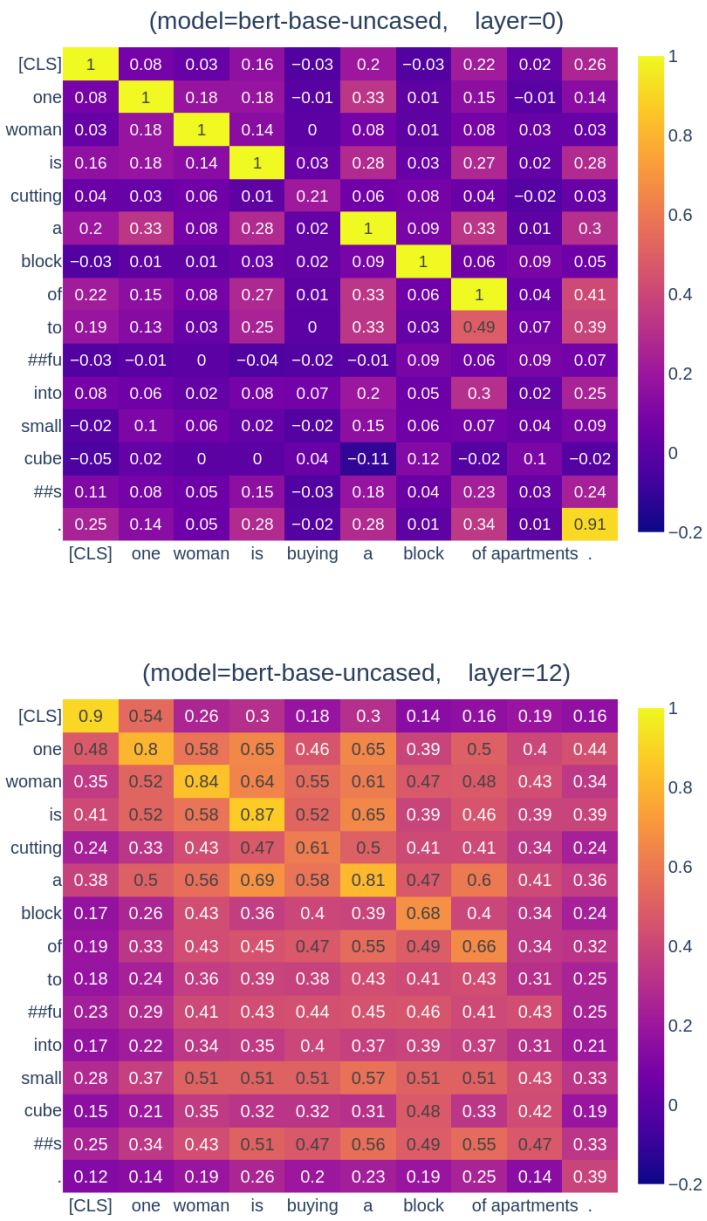

Figure 12: Cosine similarity between the token embeddings of the adversarial sentence pair from the first (top) and last (bottom) layers of bert-base-uncased model

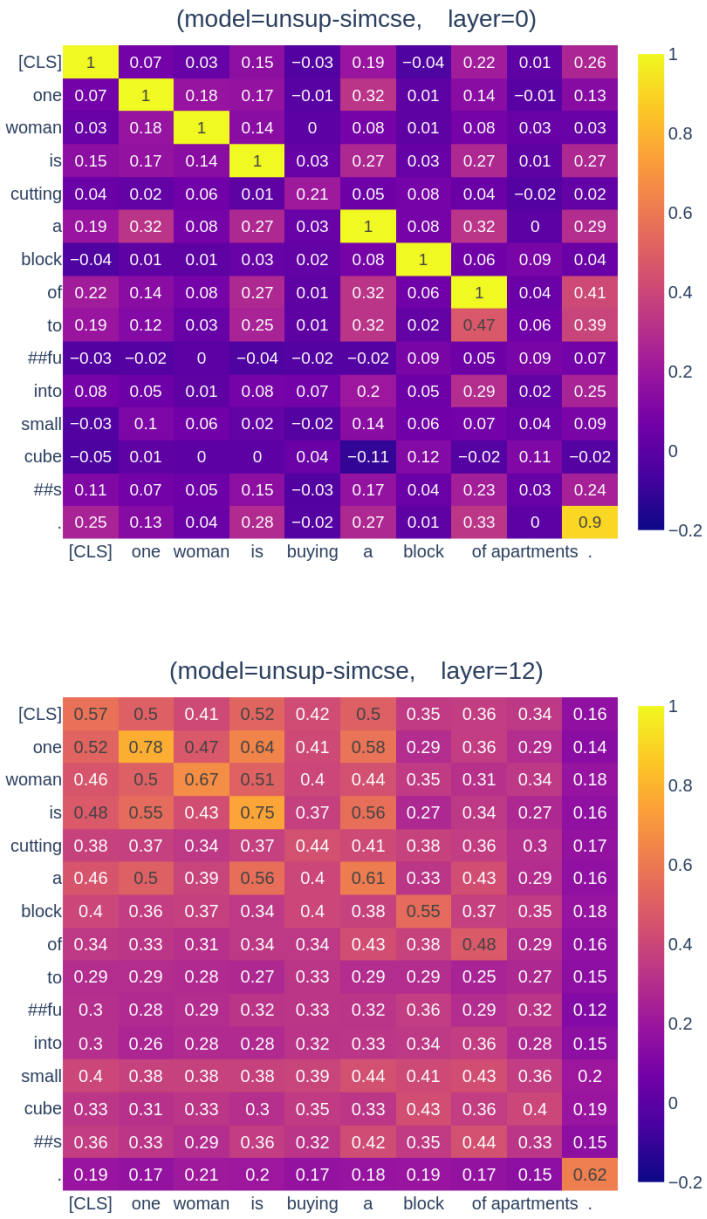

Figure 13: Cosine similarity between the token embeddings of the adversarial sentence pair from the first (top) and last (bottom) layers of the unsupervised SIMCSE model

	original	adversarial	non related
similar meaning	yes	no	no
similar syntax	no	yes	no
number of repeated words	5	7	4
bert-base-uncased (first layer)	162.2	157.17	125.5
bert-base-uncased (last layer)	180.84	159.2	136.8
unsup-simcse (first layer)	159.84	154.97	123.69
unsup-simcse (last layer)	188.77	142.52	114.19

Table 1: The pairwise similarity scores of the different pairs of sentences, for the different models and layers tested.

similarity, gives a score to the adversarial example that sits around the middle point of the range between the non related and the original scores. If we assume that the SIMCSE rated correctly the original and non related pairs with their scores corresponding with a 4 and a 0 in a scale of 0-5, the adversarial example score would be around a 2, which in STSB Cer et al. (2017) means "The two sentences are not equivalent, but share some details", were it should be a 0, that means "The two sentences are completely dissimilar". This gives us faith in that there is still a large margin for improvement in this field.

4 Conclusions

In this paper we carry out a set of experiments intended to confirm and measure known biases, and to understand their impact over anisotropy and semantic isometry in finetuned and non finetuned 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 is a correlation between the biases and the semantic isometry. These biases are present in the embedding space, making it encode information that is not semantic, like the frequency of a token or its case. This non semantic information distorts the cosine similarity, which leads to poor performance on semantic tasks.

Due to this cosine similarity distortion, biases naturally contribute to anisotropy, so there is, in fact, sometimes 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 substatial improvements over their base models, which has been

recently proved 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.

We have proposed a new similarity method that we call "pairwise similarity" aimed to alleviate the effects of biases in semantic performance. This method improves the semantic isometry of the BERT model in lower layers to a similar extent that the manual bias removal moethod does. However, this improvement is not produced in upper layers as well. As we have deduced from our experimentation, the reason for the lack of success of pairwise similarity on upper layers is the contextuality of these layers, for which it does not make sense to compare the words individually between them.

Our experimentation with pairwise similarity brought us to the conclusion that the models tested, including SIMCSE, can be fooled by repeating words and syntactic patterns. This opens a gap between the state of the art AI techniques and human performance in semantic tasks. We think that there is still a relevant margin of improvement in language understanding.

5 Future work

There are small extensions that can be done to this work, that have already been introduced in their related sections, like doing a more conclusive experiment for checking if the punctuation bias is just a subset of the frequency bias or if if has its own entity and causes. Another experiment in the line of this work would be to repeat the experimentation but separating the cased tokens in two categories: "named entities" and "beginning of sentence".

One of the techniques we used in this work was bias removal, but despite its success, we are still far away from state-of-the-art unsupervised contrastive learning models (around 0.1 below in STSB). We think that this could indicate that there are a series of biases that we still don't understand. One promising path for future 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. Maybe, we could find that the bias of some punctuation marks is also positional, as ".", or "?" are always located at the end of a sentence.

Another path that we could follow in the future is the one of studying the module of token embeddings. We have centered most of our work in measuring cosine similarity of

embeddings for checking the isotropy of a model or the semantic similarity between sentences, but we have neglected the modules of these embeddings. The modules of the token embeddings have a paramount effect on its average, which is then used as the sentence embedding. Understanding which words in which contexts have a bigger module can provide us with new insights about the representation degradation problem in transformer models.

Finally, even if our pairwise similarity metric does not work in higher layers of pretrained models, there exists the option of retraining a state of the art contrastive learning model like SIMCSE with it. It may work, or nor. That is something that we cannot know in advance, but if it worked, it would be a more explainable and controllable method for computing the similarity of two sentences than just using the cosine similarity between the average embeddings. There are other variants that could be tried here as well, like, instead of computing the alignment only between single tokens, adding also bigrams and trigrams. This could help to get the full meaning of subwords that have been splitted for being out of vocabulary or phrasal verbs, for example.

6 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 could contribute to the objective of understanding the lack of semantics in Transformer Language Models.

7 Acknowledgements

Thanks to my thesis director Víctor Fresno Fernández, that has helped me in all the stages of this work, that include, but are not limited to, hypothesis creation, experimentation design and results analysis. Thanks to Voicemod S.L., the company in which I work, for allowing me to carry out this work during my working hours. I hope that this research can contribute to the creation of future products in Voicemod.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. https://doi.org/10.18653/v1/D15-1075 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. https://openreview.net/forum?id=xYGNO86OWDH Isotropy in the Contextual Embedding Space: Clusters and Manifolds.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. https://doi.org/10.18653/v1/S17-2001 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. https://doi.org/10.18653/v1/n19-1423 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. https://aclanthology.org/2022.insights-1.1 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. https://doi.org/10.18653/v1/D19-1006 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.

http://arxiv.org/abs/2109.00725 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 Tie-Yan Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Mod-. page 14.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. https://doi.org/10.18653/v1/2021.emnlp-main.552 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. https://doi.org/10.18653/v1/2021.acl-long.72 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. http://arxiv.org/abs/2201.04337 PromptBERT: Improving BERT Sentence Embeddings with Prompts. *arXiv:2201.04337 [cs]*. ArXiv: 2201.04337.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. https://doi.org/10.18653/v1/2021.findings-acl.300 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. https://doi.org/10.18653/v1/2020.emnlp-main.733 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. https://doi.org/10.18653/v1/2021.acl-long.413 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. https://doi.org/10.3115/v1/D14-1162 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. https://doi.org/10.18653/v1/D19-1410 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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html 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. http://arxiv.org/abs/2105.11741 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. https://doi.org/10.18653/v1/2021.acl-long.402 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.