

Unsupervised Semantic Variation Prediction using the Distribution of Sibling Embeddings

Taichi Aida

Tokyo Metropolitan University
aida-taichi@ed.tmu.ac.jp

Danushka Bollegala

Amazon, University of Liverpool
danushka@liverpool.ac.uk

Abstract

Languages are dynamic entities, where the meanings associated with words constantly change with time. Detecting the semantic variation of words is an important task for various NLP applications that must make time-sensitive predictions. Existing work on semantic variation prediction have predominantly focused on comparing some form of an averaged contextualised representation of a target word computed from a given corpus. However, some of the previously associated meanings of a target word can become obsolete over time (e.g. meaning of *gay* as *happy*), while novel usages of existing words are observed (e.g. meaning of *cell* as a mobile phone). We argue that mean representations alone cannot accurately capture such semantic variations and propose a method that uses the entire cohort of the contextualised embeddings of the target word, which we refer to as the *sibling distribution*. Experimental results on SemEval-2020 Task 1 benchmark dataset for semantic variation prediction show that our method outperforms prior work that consider only the mean embeddings, and is comparable to the current state-of-the-art. Moreover, a qualitative analysis shows that our method detects important semantic changes in words that are not captured by the existing methods.¹

1 Introduction

The meaning of words evolves over time, and even in everyday life, technological innovations and cultural aspects can cause a word to have a different meaning than in the past. For example, the meaning of the word *gay* has completely changed from *happy* to *homosexual* (Figure 1a), and *cell* has added *cell phone* to its previous meanings of *prison* and *biology* (Figure 1b). In the semantic change detection task, the goal is to detect the words whose meanings have changed across time-specific corpora (Kutuzov et al., 2018; Tahmasebi et al., 2021).

¹Source code is available at <https://github.com/a1da4/svp-gauss>.

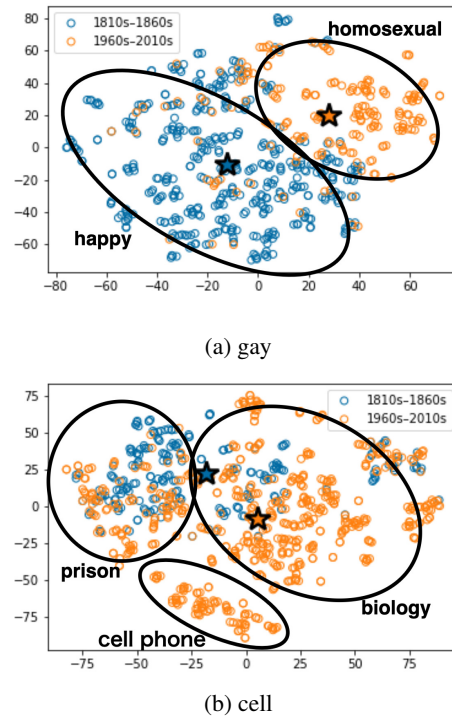


Figure 1: t-SNE projections of BERT token vectors (dotted) in two time periods and the average vector (starred) for each period. (a) the word *gay* has lost its original meaning related to *happy* and is now used to mean *homosexual*, resulting in a significant shift in its distribution. (b) the word *cell* is now also used to mean *cell phone*, while retaining the meaning of *prison* or *biology*, widening the distribution but not significantly changing the mean vector.

As illustrated in Figure 1, we can identify two types of semantic changes associated with words – (a) the word *gay* obtains a new meaning by **replacing** its past meaning (Figure 1a), whereas (b) the word *cell* obtains a new meaning, while **preserving** its past meanings (Figure 1b). On the other hand, much prior work have resort to a scheme where they first individually represent the meaning of a target word in a given time-specific corpora using a single embedding, such as the mean of the non-contextualised (Kim et al., 2014; Kulka-rni et al., 2015; Hamilton et al., 2016; Yao et al.,

2018; Dubossarsky et al., 2019; Aida et al., 2021) or contextualised (Martinc et al., 2020; Beck, 2020; Kutuzov and Giulianelli, 2020; Rosin et al., 2022; Rosin and Radinsky, 2022) embeddings of the target word taken over all of its occurring contexts in the corpus. Next, various distance measures are used to compare those embeddings to quantify the semantic variation of the target word across corpora. However, as seen from Figure 1, using the mean embedding of a target word alone for predicting semantic variations of words can be misleading especially when the variance of the embedding distribution is large.

To address the above-mentioned limitations, we use the distribution of contextualised embeddings of a target word w in all of its occurrence contexts $\mathcal{S}(w)$ in a given corpus, which we refer to as the *sibling distribution* (Zhou et al., 2022) of w . We then approximate the sibling distribution of a word using a multivariate Gaussian, which has shown to accurately capture the uncertainty in word embedding spaces (Vilnis and McCallum, 2015; Iwamoto and Yukawa, 2020; Yüksel et al., 2021). We can then use a broad range of distance and divergence measures defined over Gaussian distributions to quantify the semantic variation of a target word across multiple time-specific corpora.

Experimental results on SemEval-2020 Task 1 benchmark dataset show that our proposed method outperforms several prior methods, and achieves comparable performance to the current state-of-the-art (SoTA) (Rosin and Radinsky, 2022). More importantly, our proposal to model both the mean and variance of sibling embeddings consistently outperforms methods that use only the mean contextualised embedding from the same Masked Language Model (MLM) (Rosin and Radinsky, 2022). Moreover, for computational convenience, prior work had assumed the covariance matrix of sibling embeddings to be diagonal (Iwamoto and Yukawa, 2020; Yüksel et al., 2021), but we show that further performance improvements can be obtained by using the full covariance matrix.

2 Related Work

Historically, the diachronic semantic changes of words have been studied by linguists (Tahmasebi et al., 2021), which has also received much attention lately within the NLP community. Automatic detection of words whose meanings change over time has provided important insights for diverse

fields such as linguistics, lexicology, sociology, and information retrieval (IR) (Traugott and Dasher, 2001; Cook and Stevenson, 2010; Michel et al., 2011; Kutuzov et al., 2018). For example, in IR one must know the seasonal association of keywords used in user queries to provide relevant results pertaining to a particular time period. Moreover, it has been shown that the performance of publicly available pretrained foundation models (Bommasani et al., 2021) declines over time when applied to emerging data (Loureiro et al., 2022; Lazaridou et al., 2021) because they are trained using a static snapshot. Su et al. (2022) showed that the temporal generalisation of foundation models is closely related to their ability to detect semantic variations of words.

Semantic change detection is modelled in the literature as an unsupervised task of detecting words whose meanings change between two given time-specific corpora (Kutuzov et al., 2018; Tahmasebi et al., 2021). In recent years, several shared tasks have been held (Schlechtweg et al., 2020; Basile et al., 2020; Kutuzov and Pivovarova, 2021), where participants are required to predict the degree or presence of semantic changes for a given target word between two given corpora sampled from different time periods. For this purpose, much prior work have used non-contextualised or contextualised word embeddings to represent the meaning of the target word in each corpus. Unlike non-contextualised word embeddings, which represent a word by the same vector in all of its contexts, contextualised word embeddings represent the same target word with different vectors in different contexts. Various methods have been proposed to map vector spaces from different time periods, such as initialisation (Kim et al., 2014), alignment (Kulkarni et al., 2015; Hamilton et al., 2016), and joint learning (Yao et al., 2018; Dubossarsky et al., 2019; Aida et al., 2021).

The existing methods that have been proposed for the semantic variation detection of words can be broadly categorised into two groups: (a) methods that compare word/context clusters (Hu et al., 2019; Giulianelli et al., 2020; Montariol et al., 2021), and (b) methods that compare embeddings of the target words computed from different corpora sampled at different time periods (Martinc et al., 2020; Beck, 2020; Kutuzov and Giulianelli, 2020; Rosin et al., 2022). Recently, it has been reported that adding time-specific attention mechanisms (Rosin

and Radinsky, 2022) achieves SoTA performance. However, this model requires additional training of the entire MLM including the time-specific mechanisms, which is computationally costly for large-scale MLMs.

Despite the recent success of using word embeddings for the semantic change detection task, many of these methods struggle to detect meaning changes of words which have a wide range of usages because they use only the mean embedding to represent a target word (Kutuzov et al., 2022). Although methods that use point estimates in the embedding space, such as using non-contextualised word embeddings or comparing the average of contextualised word embeddings, are able to detect semantic variations that result in a loss of a prior meaning (e.g. *gay* in Figure 1a), they are inadequate when detecting semantic variations due to novel usages of words, while preserving their former meanings (e.g. *cell* in Figure 1b).

To alleviate this problem, some studies have used Gaussian Embeddings (Vilnis and McCallum, 2015) for semantic change detection (Iwamoto and Yukawa, 2020; Yüksel et al., 2021). They used the mean and the diagonal approximation of the covariance matrix computed using non-contextualised word embeddings. However, as argued previously, contextualised embeddings provide useful clues regarding the meaning of a word as used in a context. Therefore, in our proposed method, we consider the entire cohort of contextualised word embeddings of a target word taken across all of its occurring contexts (i.e. siblings) obtained from an MLM. As confirmed later by the evaluations presented in § 4.4, our proposed method consistently outperforms the methods proposed by Iwamoto and Yukawa (2020) and Yüksel et al. (2021) that use non-contextualised embeddings.

3 Semantic Variation Prediction

Let us consider a target word w that occurs in two given corpora C_1 and C_2 . For example, C_1 and C_2 could have been sampled at two distinct time slots, respectively T_1 and T_2 , reflecting any *temporal* semantic variations of words, or alternatively sampled at similar periods in time but from distinct domains (e.g. *biology* vs. *law*) expressing semantic variations of words due to the differences in the *domains*. Our goal in this paper is to propose a method that can accurately predict whether w is used in the same meaning in both C_1 and C_2 (i.e.

w is semantically invariant across the two corpora) or otherwise (i.e. its meaning is different in the two corpora). Although we consider two corpora in the subsequent description for simplicity of the disposition, our proposed method can be easily extended to measure the semantic variation of a word over multiple corpora.

According to the distributional hypothesis (Firth, 1957), the context in which a word occurs provides useful clues regarding its meaning. Contextualised word embeddings such as the ones produced by MLMs have shown to concisely and accurately encode contextual information related to a target word in a given context. For example, Zhou and Bollegala (2021) showed that contextualised word embeddings can be used to induce word-sense embeddings that represent the distinct senses of an ambiguous word with different vectors. Inspired by such prior work using contextualised word embeddings as a proxy for accessing contextual information related to a target word, we propose a method to detect the semantic variations of a target word using its multiple occurrences in a corpus.

To describe our proposed method in detail, let us denote the set of contexts containing w in corpus C_i by $\mathcal{S}(w, C_i)$. The scope of the context of w could be limited to a predefined fixed token window or extended to the entire sentence containing w as we do in our experiments. Let us denote the contextualised (token) embedding of w in a context $s \in \mathcal{S}(w, C_i)$ produced by an MLM M by $\mathbf{f}_M(w, s) \in \mathbb{R}^d$, where d is the dimensionality of the token embeddings produced by M . Following the terminology introduced by Zhou et al. (2022), we refer to type embedding $\mathbf{f}_M(w, s)$ as the *sibling* embeddings of w in context s . The number of siblings of w in C_i is denoted by $N_i^w = |\mathcal{S}(w, C_i)|$. Moreover, let the set of sibling embeddings of w created from its occurrences in C_i to be $\mathcal{D}(w, C_i) = \{\mathbf{f}_M(w, s) | s \in \mathcal{S}(w, C_i)\}$. As we later see, the distribution of sibling embeddings of a word w encodes information about the usage of w in a corpus, which is useful for predicting any semantic variations of w across different corpora.

We can obtain a context-independent embedding, $\boldsymbol{\mu}_i^w \in \mathbb{R}^d$ for w by averaging all of its sibling embeddings over the contexts as given by (1).

$$\boldsymbol{\mu}_i^w = \frac{1}{N_i^w} \sum_{s \in \mathcal{S}(w, C_i)} \mathbf{f}_M(w, s) \quad (1)$$

Although much prior work has used $\boldsymbol{\mu}_i^w$ as a proxy

for the usage of w in C_i for numerous tasks such as studying the properties of contextualised embeddings (Ethayarajh, 2019) and predicting semantic variation of words (Martinc et al., 2020; Beck, 2020; Kutuzov and Giulianelli, 2020; Rosin et al., 2022; Rosin and Radinsky, 2022), the mean of the sibling embedding distribution is insensitive to the rare yet important usages of the target word. In particular, when the sibling embedding distribution is not uniformly distributed around its mean, the mean embedding can be misleading as a representation of the distribution. To overcome this limitation, in addition to $\boldsymbol{\mu}_i^w$, we also use the covariance matrix $\mathbf{V}_i^w \in \mathbb{R}^{d \times d}$ computed from the sibling embedding distribution of w as defined by (2).

$$\mathbf{V}_i^w = \frac{1}{N_i^w(N_i^w - 1)} \sum_{s \in \mathcal{S}(w, C_i)} \mathbf{f}_{M(w, s)} \mathbf{f}_{M(w, s)}^\top \quad (2)$$

We approximate the distribution of sibling embeddings of w using a Gaussian, $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{V}_i^w)$ with mean and variance given respectively by (1) and (2). Gaussian distribution is the maximum entropy distribution over the real values given a finite mean and covariance and no further information (Jaynes, 2003). Moreover, by approximating the sibling distribution as a Gaussian, we can use a broad range of distance and divergence measures for quantifying the semantic variation of w across corpora. In the field of information theory, MLMs have been shown to store the information of a given sentence in a vector (Pimentel et al., 2020). There is a strong correlation between the word frequency N_i^w and the rank of its covariance matrix \mathbf{V}_i^w (Figure 2 in Appendix A), which indicates that covariance matrix also retains important information regarding sibling embedding distribution. This observation further supports our proposal to represent target words by $\boldsymbol{\mu}_i^w$ and \mathbf{V}_i^w .

3.1 Quantifying Semantic Variations

Given a target word w , following the method described above, we represent w in C_1 and C_2 respectively by the two Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1^w, \mathbf{V}_1^w)$ and $\mathcal{N}(\boldsymbol{\mu}_2^w, \mathbf{V}_2^w)$. We can then compute a *semantic variation score* for w that indicates how likely the meaning of w has changed from C_1 to C_2 by using different distance (or divergence) measures to quantify the differences between two Gaussians. For this purpose, we use two types of measures.

Divergence measures quantify the divergence between two distributions. We use two divergence measures in our experiments: Kullback-Liebler (KL) divergence and Jeffrey’s divergence. Given that we approximate sibling distribution of w in a corpus by a Gaussian, we can analytically compute both KL and Jeffrey’s divergence measures using $\boldsymbol{\mu}_1^w, \boldsymbol{\mu}_2^w, \mathbf{V}_1^w$ and \mathbf{V}_2^w in closed-form formulas (Appendix B).

Distance measures are defined between two points in the sibling embedding space. We use the seven distance measures: Bray-Curtis, Canberra, Chebyshev, City Block, Correlation, Cosine, and Euclidean. The definitions of the distance measures used in this paper are provided in Appendix C. Given a distance measure $\psi(\mathbf{w}_1, \mathbf{w}_2)$ that takes two d -dimensional sibling embeddings of w , each computed from contexts selected respectively from C_1 and C_2 and returns a nonzero real number indicating the distance between \mathbf{w}_1 and \mathbf{w}_2 , we compute the semantic variation score, $\text{score}(w)$, of w between C_1 and C_2 as the average distance over all pairwise comparisons between the sibling embeddings as given by (3).

$$\text{score}(w) = \frac{1}{N_1^w N_2^w} \sum_{\substack{\mathbf{w}_1 \in \mathcal{D}(w, C_1) \\ \mathbf{w}_2 \in \mathcal{D}(w, C_2)}} \psi(\mathbf{w}_1, \mathbf{w}_2) \quad (3)$$

The number of occurrences of some target words w can be significantly different between C_1 and C_2 , which can make the computation of (3) biased towards the corpus with more contexts for w . To overcome this issue, instead of using sibling embeddings of w computed from actual occurrence contexts of w , we sample equal numbers of sibling embeddings from $\mathcal{N}(\boldsymbol{\mu}_1^w, \mathbf{V}_1^w)$ and $\mathcal{N}(\boldsymbol{\mu}_2^w, \mathbf{V}_2^w)$. Samples can be drawn efficiently from a multidimensional Gaussian by first drawing samples from a standard normal distribution (i.e. with zero mean and unit variance) and subsequently applying an affine transformation parametrised by the $\boldsymbol{\mu}_i^w$ and \mathbf{V}_i^w of the associated sibling distribution.

4 Experiments

4.1 Data and Metric

We use the SemEval-2020 Task 1 English dataset² (Schlechtweg et al., 2020) to evaluate the performance in detecting words whose meanings change between time periods. This task includes

²It is licensed under a Creative Commons Attribution 4.0 International License.

Time Period	#Sentences	#Tokens	#Types
1810s–1860s	254k	6.5M	87k
1960s–2010s	354k	6.7M	150k

Table 1: Statistics of the SemEval-2020 Task 1 English dataset (Schlechtweg et al., 2020).

two subtasks, classification and ranking. In the classification task, the words in the evaluation set must be classified as to whether they have semantically changed over time or otherwise. Classification accuracy is used as the evaluation metric for this task. On the other hand, in the ranking task, the words in the evaluation set must be sorted according to the degree of semantic change. Spearman’s rank correlation coefficient between the human-rated gold scores and the induced ranking scores is used as the evaluation metric for this task. In this study, the evaluation is conducted on the ranking task using English data. We do not perform the classification task because no validation set is available for tuning a classification threshold.

Statistics of the data used in our experiments are in Table 1. This data includes two corpora from different centuries extracted from CCOHA (Alatrash et al., 2020). Let us denote the early 1800s and late 1900s to early 2000s corpora respectively by C_1 and C_2 . The test set has 37 target words that are selected for indicating whether they have undergone a semantic change between the two time periods. These words are annotated indicating whether their meaning has changed over time and the degree of their semantic change.

4.2 Setup

We use two types of BERT-base models as the MLM in our experiments: a publicly available pre-trained model³ (MLM_{pre}) and a fine-tuned model (MLM_{temp}) from MLM_{pre} (Rosin et al., 2022). The base model consists of 12 layers, which we use in two different configurations: (a) we use the last layer ($\text{MLM}_{pre|temp,last}$), and (b) the mean-pool over the last four layers ($\text{MLM}_{pre|temp,four}$), which has shown good performance across languages following Laicher et al. (2021). Rosin and Radinsky (2022) recommend using the mean pooling over all (12) hidden layers. However, we found no statistically significant differences between the mean-pool over all layers vs. the last four layers in our

³<https://huggingface.co/bert-base-uncased>

preliminary experiments.

In the prediction of the degree of semantic change for a given word, the set of sibling embeddings for each time period $\mathcal{D}(w, C_1)$ and $\mathcal{D}(w, C_2)$ is acquired from all occurrences in each corpus using the MLM described above, and the distributions across time periods $\mathcal{N}(\boldsymbol{\mu}_1^w, \mathbf{V}_1^w)$ and $\mathcal{N}(\boldsymbol{\mu}_2^w, \mathbf{V}_2^w)$ are compared. For calculating the seven distance measures, we sample 1,000 sibling embeddings from each sibling distribution. We use the covariance matrix of the sibling embedding, which defines the distribution, only for the diagonal components ($diag(cov)$) in the divergence measures,⁴ and both diagonal and full components ($full(cov)$) in the distance measures. Previous studies assume that the covariance matrix is diagonal ($diag(cov)$) (Iwamoto and Yukawa, 2020; Yüksel et al., 2021). This assumption increases computational efficiency compared to $full(cov)$, at the expense of losing information on the non-diagonal elements. In our settings, representation of a sibling distribution $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{V}_i^w)$ in $diag(cov)$ or $full(cov)$ requires $2d$ or $d(1 + d)$ parameters, respectively.

4.3 Result

We show the results of the proposed method under various conditions in Table 2 and Table 3. As reported in previous studies (Rosin et al., 2022; Rosin and Radinsky, 2022), we find that the fine-tuned model (MLM_{temp}) achieves high performance in all settings. Moreover, for the hidden layers, we have confirmed that our method, by using the last four layers ($\text{MLM}_{pre|temp,four}$), yields even higher correlations than using only the last layer ($\text{MLM}_{pre|temp,last}$).

Prediction measures. Our method allows us to try a variety of measures. In the $diag(cov)$ setting, we try two divergences and seven distance measures. Comparing within divergence measures, Table 2 shows that $KL(C_1||C_2)$ achieves high performance in all MLM conditions. This result means that many existing words acquire novel meanings. On the other hand, comparing the distance measures, we find that Canberra and Chebyshev outperform the commonly used cosine distance in MLM_{temp} (Table 2 and Table 3). Since the cosine distance makes underestimations in MLMs (Zhou

⁴In the above two divergences, it is necessary to calculate the inverse of the covariance matrix, but in the case of full components, it is often impossible to calculate the inverse matrix because it is not regular.

Measure	Model			
	MLM _{pre,last}	MLM _{pre,four}	MLM _{temp,last}	MLM _{temp,four}
KL($C_1 C_2$)	0.075	0.130	0.414	0.431
KL($C_2 C_1$)	0.100	0.117	0.361	0.411
Jeff($C_1 C_2$)	0.090	0.129	0.391	0.409
Bray-Curtis	0.217	0.241	0.464	0.480
Canberra	0.192	0.251	0.455	0.517
Chebyshev	0.154	0.166	0.517	0.478
City Block	0.198	0.140	0.461	0.459
Correlation	0.191	0.266	0.480	0.463
Cosine	0.190	0.270	0.478	0.480
Euclidean	0.198	0.249	0.473	0.474

Table 2: Results of two divergences and seven distance functions under various MLM conditions with the proposed method using *diag(cov)*. The best performance in each MLM condition is shown in **bold**. C_1 and C_2 refer to the early 1800s and late 1900s to early 2000s corpora, respectively. We report two types of KL divergence because of its asymmetric nature. Unlike KL divergence, Jeffrey’s divergence is symmetric, and we report just one result.

Measure	Model			
	MLM _{pre,last}	MLM _{pre,four}	MLM _{temp,last}	MLM _{temp,four}
Bray-Curtis	0.219	0.263	0.460	0.467
Canberra	0.195	0.246	0.502	0.489
Chebyshev	0.145	0.132	0.529	0.451
City Block	0.192	0.248	0.414	0.452
Correlation	0.181	0.286	0.481	0.468
Cosine	0.189	0.272	0.479	0.454
Euclidean	0.204	0.231	0.454	0.457

Table 3: Results of two divergences and seven distance functions under various MLM conditions with the proposed method using *full(cov)*. The best performance in each MLM condition is shown in **bold**.

et al., 2022), this result suggests that it is better to calculate the absolute distance per dimension as in Canberra and Chebyshev.

Components of the covariance matrices. When applying the distance measures, the vectors can be extracted from the full or diagonal covariance matrix. From Table 3 we see that using all components of the covariance matrix (*full(cov)*) further improves performance obtaining a correlation coefficient of 0.529 (MLM_{temp,last}, *full(cov)*, Chebyshev). Previous studies had assumed that the covariance matrix is diagonal for computational convenience (Iwamoto and Yukawa, 2020; Yüksel et al., 2021). However, as our results show, further performance improvements can be obtained by considering all components of the covariance matrix. Here onwards, we will refer to the best setting (i.e. MLM_{temp,last}, *full(cov)*, Chebyshev) as the **Proposed** method.

4.4 Comparisons against Prior work

In this section, we compare our proposed method against related prior work. We do not re-implement or re-run those methods, but instead compare using the published results from the original papers.

Word2Gauss_{light} (Iwamoto and Yukawa, 2020):

They apply Gaussian Embeddings (Vilnis and McCallum, 2015) based architecture in each time period. For each word, they define a computationally lightweight Gaussian embedding as follows: the mean vector is the vector of the word2vec learned by the initialization method (Kim et al., 2014), and the covariance matrix is the diagonal matrix, uniformly weighted by frequency. They calculate the KL divergence of the Gaussian embeddings for the semantic variation prediction.

Word2Gauss (Yüksel et al., 2021): They apply

pure Gaussian Embeddings (Vilnis and McCallum, 2015). For a given word, the mean vector and the covariance matrix of the Gaussian Embedding are trained using the inner-product with the positive examples and the KL divergence with the negative examples. For computational convenience and to reduce the number of parameters, they use a diagonal covariance matrix. After training separate word embedding models for each time period, the mean vectors are aligned between time periods using a rotation matrix (Hamilton et al., 2016), and predictions are made using cosine distance or Jeffrey’s divergence. They have reported the cosine distance as the best metric.

MLM_{temp} (Rosin et al., 2022): They fine-tuned the published BERT model to specific time periods. To adapt to specific time periods, they insert a special token indicating the time period at the beginning of the sentence in the target corpus, and fine-tuned on the corpora available for each time period. They use two measures for prediction: (a) the distance between the predicted probability of the target word in the sentence at each time period, and (b) the cosine distance of the average token vector at each time period. Their results report that the cosine distance is the best metric (MLM_{temp}, Cosine). However, Kutuzov and Giulianelli (2020) have shown that the average pairwise cosine distance (3) is better than the cosine distance between average sibling embeddings. Based on this result, we only run this setting that MLM_{temp} model with the average pairwise cosine distance (MLM_{temp}, APD).

MLM_{pre} w/ Temp. Att. (Rosin and Radinsky, 2022): They propose a time-specific attention mechanism to adapt MLMs to specific time periods. They add time-specific vectors and an attention weight matrix to the published BERT as trainable parameters and perform additional training on the target corpora. During prediction, they use the cosine distance following Rosin et al. (2022).

MLM_{temp} w/ Temp. Att. (Rosin and Radinsky, 2022): It is the combination of the above two methods (MLM_{temp} and MLM_{pre} w/ Temp. Att.), which is considered as the current SoTA model for semantic variation prediction. They

Model	Spearman
Word2Gauss _{light}	0.358
Word2Gauss	0.399
MLM _{temp} , Cosine	0.467
MLM _{temp} , APD	0.479
MLM _{pre} w/ Temp. Att.	0.520
MLM _{temp} w/ Temp. Att.	0.548
Proposed	<u>0.529</u>

Table 4: Comparison against prior work including SoTA. In our method, we report the top three results and all of the cosine distance results. The best performance is shown in **bold**, and the second best is shown in underlined.

add time-specific special tokens to the beginning of each sentence in the target corpus, and conduct additional training on the publicly available BERT model with the time-specific attention mechanism. They also use the cosine distance as used by Rosin et al. (2022).

Experimental results are summarised in Table 4. This result shows that our proposed method achieves the second best performance compared to prior work. We can see that the contextualised mean embeddings based method (MLM_{temp}) outperforms the non-contextualised distribution based methods (Word2Gauss_{light} and Word2Gauss), and further improvement can be obtained by adding the time-specific attention mechanisms (MLM_{pre} w/ Temp. Att. and MLM_{temp} w/ Temp. Att.). Moreover, the contextualised distribution based approach (Proposed) can yield performance improvement similar to adding time-specific attention mechanisms. We will discuss the detailed analyses as follows.

Comparison within the base model (MLM_{temp}).

Since our method is based on MLM_{temp}, we compare performance within MLM_{temp}. As in the previous work (Rosin et al., 2022), we discuss the results when using the cosine distance. Table 4 shows that the average pairwise cosine distance (MLM_{temp}, APD) outperforms the cosine distance between average sibling embeddings (MLM_{temp}, Cosine). Moreover, from Table 2 and Table 3, we can see that our distribution based method outperforms the previous method using only the mean embeddings (0.467 in Table 4) in most settings (0.478 by MLM_{temp,last}, diag(cov), 0.480 by MLM_{temp,four}, diag(cov), and 0.479 by

$\text{MLM}_{temp, last, full(cov)}$). This result indicates the importance of considering not only the mean but also the variance of the sibling embeddings.

Comparison against SoTA. Although our proposed method and the SoTA MLM_{temp} **w/ Temp. Att.** are based on the same model MLM_{temp} , their configurations are significantly different. Specifically, MLM_{temp} **w/ Temp. Att.** adds a time-specific attention mechanism to the model and learns its parameters with additional training, whereas our proposed method uses only MLM_{temp} and thus does *not* require additional parameters or training. Although according to Table 4, MLM_{temp} **w/ Temp. Att.** reports a correlation of 0.548 and marginally outperforms the Proposed method, which obtains a correlation of 0.529, we find no statistically significant difference between those two methods.⁵

4.5 Ablation Study

We conduct an ablation study to understand the importance of (i) predicting semantic variation with sibling distributions $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{V}_i^w)$, and (ii) constructing sibling distributions from the mean $\boldsymbol{\mu}_i^w$ and covariance \mathbf{V}_i^w of sibling embeddings. Based on our best setting **Proposed** ($\text{MLM}_{temp, last, full(cov)}$, Chebyshev), we define two variants: (i) predicting semantic variation score using mean vectors $\boldsymbol{\mu}_1^w$ and $\boldsymbol{\mu}_2^w$ only as previous studies, and (ii) constructing a sibling distribution with the identity matrix $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{I})$ instead of the covariance matrix \mathbf{V}_i^w . In the SemEval-2020 Task 1 English evaluation set, the existence of a semantic change (binary judgement) and its degree (continuous judgement) are provided. Therefore, due to the limited space, we analyse the top eight semantically changed words with the highest degrees of semantic changes and the bottom eight semantically stable words with the lowest degrees of semantic change.

From Table 5, we see that our distribution-based variants ($\mathbf{V}_i^w = \mathbf{I}$ and Proposed) eliminate overestimation or underestimation problems in using mean vectors only (w/o \mathbf{V}_i^w). The variant w/o \mathbf{V}_i^w correctly detects words *plane* and *graft* that have changed meaning significantly between time periods. However, this variant also reports underestimation (*stab* and *bit*) and overestimation (*contemplation* and *chairman*) in other words, whose meanings are changed/stable but the mean vec-

⁵To measure the statistical significance, we use the Fisher transformation (Fisher, 1992).

Word	Gold		w/o \mathbf{V}_i^w rank	$\mathbf{V}_i^w = \mathbf{I}$ rank	Proposed rank
	rank	Δ			
plane	1	✓	3	18	15
tip	2	✓	7	9	7
prop	3	✓	16	1	4
graft	4	✓	2	36	36
record	5	✓	15	12	14
stab	7	✓	31	10	11
bit	9	✓	27	11	9
head	10	✓	23	28	28
<hr/>					
multitude	30	✗	24	35	35
savage	31	✗	20	26	26
contemplation	32	✗	1	37	37
tree	33	✗	33	31	30
relationship	34	✗	26	34	34
fiction	35	✗	21	29	29
chairman	36	✗	5	32	33
risk	37	✗	10	19	21
<hr/>					
Spearman	1.000		0.070	0.503	0.529

Table 5: Ablation study on the top-8 semantically changed ($\Delta = \checkmark$) words with the highest degree of semantic change and the bottom-8 stable words ($\Delta = \times$) with the lowest degree of semantic change. w/o \mathbf{V}_i^w predicts using mean vectors $\boldsymbol{\mu}_1^w$ and $\boldsymbol{\mu}_2^w$ directly, whereas $\mathbf{V}_i^w = \mathbf{I}$ samples sibling embeddings from a Gaussian with the identity variance (i.e. $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{I})$) instead of $\mathcal{N}(\boldsymbol{\mu}_i^w, \mathbf{V}_i^w)$.

tors are changed little/significantly. This is because it makes predictions based only on the mean of sibling embeddings. On the other side, the distribution-based variants ($\mathbf{V}_i^w = \mathbf{I}$ and Proposed) can appropriately rank semantically changed words ($\Delta = \checkmark$) that have small changes in mean vectors (*stab* and *bit*), and stable words ($\Delta = \times$) that have large changes in mean vectors (*contemplation* and *chairman*).⁶ Moreover, we find that even with the distribution-based variants, using covariance matrices \mathbf{V}_i^w computed from sibling embeddings yields even better performance than identity matrices ($\mathbf{V}_i^w = \mathbf{I}$). This result further verifies our hypothesis that considering the mean and the variance of the sibling embeddings is important for semantic change detection tasks.

5 Conclusion

We proposed a method to detect semantic variations of words using sibling embeddings. Experimental results on SemEval-2020 Task 1 English

⁶The distribution-based methods fail to detect highly ambiguous words with distinct word senses (*plane* and *graft*). However, the proposed method approximates the distribution of embeddings for a word using a “single” Gaussian. We believe by using a mixture of Gaussian this issue can be resolved.

dataset show that the proposed method consistently outperforms methods that use only the mean embedding vectors, and reports results comparable to the current SoTA. Furthermore, a qualitative analysis shows that the proposed method correctly detects semantic variation of words, which are either over/underestimated by the existing methods.

6 Limitations

Language-related limitations. For the ease of the analysis, we conducted experiments using only the English dataset in this study. Although our proposed method can be applied to any language, its performance must be evaluated on languages other than English. For example, the SemEval-2020 Task 1 dataset includes Latin, German, and Swedish language datasets, in addition to English, and can be used for this purpose. In particular, our proposed method requires only pretrained MLMs and does not require additional training data for the target languages, which makes it easily scalable to many languages.

Availability of MLMs for the target language. Experimental results show that the quality of the MLM is an important factor determining the performance of the proposed method. For example, the proposed method reports good performance with vanilla BERT model in Table 2 but further gains in performance can be obtained with the fine-tuned BERT model on masked time stamps. However, since our method assumes the availability of pretrained MLMs, a problem arises when trying to adapt our method to minor languages where no pretrained MLMs are available. This limitation could be mitigated to an extent by using multilingual MLMs. For example, Arefyev and Zhikov (2020) demonstrated that satisfactory levels of accuracies can be obtained for semantic change detection by using multilingual MLMs. Our proposed method can further benefit from the fact that new and larger MLMs are being publicly released for many languages in the NLP community.

7 Ethical Considerations

In this paper, we proposed a distribution based method using publicly available MLMs, and evaluated with the SemEval-2020 Task 1 English data. Although we have not published any datasets or models, Basta et al. (2019) shows that pretrained MLMs encode and even amplify unfair social biases such as gender or racial biases. Given that we

obtain sibling distributions from such potentially socially biased MLMs, we must further evaluate the sensitivity of our method for such undesirable social biases.

Acknowledgements

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2139. Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

References

- Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. *A comprehensive analysis of PMI-based models for measuring semantic differences*. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 21–31, Shanghai, China. Association for Computational Linguistics.
- Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. *CCOHA: Clean corpus of historical American English*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.
- Nikolay Arefyev and Vasily Zhikov. 2020. *BOS at SemEval-2020 task 1: Word sense induction via lexical substitution for lexical semantic change detection*. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 171–179, Barcelona (online). International Committee for Computational Linguistics.
- Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. *Diacrita @ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task*. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*. CEUR Workshop Proceedings (CEUR-WS.org). Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2020 ; Conference date: 17-12-2020.
- Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. *Evaluating the underlying gender bias in contextualized word embeddings*. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

- Christin Beck. 2020. [DiaSense at SemEval-2020 task 1: Modeling sense change via pre-trained BERT embeddings](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58, Barcelona (online). International Committee for Computational Linguistics.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramér, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. *On the Opportunities and Risks of Foundation Models*.
- Paul Cook and Suzanne Stevenson. 2010. [Automatically identifying changes in the semantic orientation of words](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. [Time-out: Temporal referencing for robust modeling of lexical semantic change](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. 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.
- John R. Firth. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis*, pages 1 – 32.
- R. A. Fisher. 1992. *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY.
- Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. [Analysing lexical semantic change with contextualised word representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Renfen Hu, Shen Li, and Shichen Liang. 2019. [Diachronic sense modeling with deep contextualized word embeddings: An ecological view](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.
- Ran Iwamoto and Masahiro Yukawa. 2020. [RIJP at SemEval-2020 task 1: Gaussian-based embeddings for semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 98–104, Barcelona (online). International Committee for Computational Linguistics.
- E. T. Jaynes. 2003. *Probability Theory*. Cambridge University Press.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [Temporal analysis of language through neural language models](#). In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW 2015*, pages 625–635.
- Andrei Kutuzov, Erik Velldal, and Lilja Ovrelid. 2022. [Contextualized embeddings for semantic change detection: Lessons learned](#). *Northern European Journal of Language Technology*, 8.

- Andrey Kutuzov and Mario Giulianelli. 2020. [UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.
- Andrey Kutuzov, Lilja Ovreliid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Andrey Kutuzov and Lidia Pivovarov. 2021. [RuShiftEval: a shared task on semantic shift detection for Russian](#). In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.
- Severin Laicher, Sinan Kurtuyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. [Explaining and improving BERT performance on lexical semantic change detection](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomáš Kočický, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. [Mind the gap: Assessing temporal generalization in neural language models](#). In *Advances in Neural Information Processing Systems*.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. [TimeLMs: Diachronic language models from Twitter](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.
- Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. [Leveraging contextual embeddings for detecting diachronic semantic shift](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, null null, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. [Quantitative analysis of culture using millions of digitized books](#). *Science*, 331(6014):176–182.
- Syrielle Montariol, Matej Martinc, and Lidia Pivovarov. 2021. [Scalable and interpretable semantic change detection](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. [Time masking for temporal language models](#). In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM ’22*, pages 833–841, New York, NY, USA. Association for Computing Machinery.
- Guy D. Rosin and Kira Radinsky. 2022. [Temporal attention for language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [SemEval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.
- Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. [Improving temporal generalization of pre-trained language models with lexical semantic change](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nina Tahmasebi, Lars Borina, and Adam Jatowtb. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6:1.
- Elizabeth Closs Traugott and Richard B. Dasher. 2001. [Prior and current work on semantic change](#), Cambridge Studies in Linguistics, page 51–104. Cambridge University Press.
- Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.
- Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. [Dynamic word embeddings for evolving semantic discovery](#). In *WSDM 2018*, page 673–681.
- Arda Yüksel, Berke Uğurlu, and Aykut Koç. 2021. [Semantic change detection with gaussian word embeddings](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3349–3361.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

Yi Zhou and Danushka Bollegala. 2021. Learning sense-specific static embeddings using contextualised word embeddings as a proxy. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 493–502, Shanghai, China. Association for Computational Linguistics.

Yi Zhou and Danushka Bollegala. 2022. On the curious case of 12 norm of sense embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2593–2602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Information of Sibling Distribution

In the semantic variation prediction, prior work have applied the mean embeddings μ_i^w of sibling distribution $\mathcal{D}(w, C_i)$ for each word w . However, since these methods compress multiple vectors of $\mathcal{D}(w, C_i)$ into a single vector μ_i^w , there is a risk of losing the information contained in each vector (Pimentel et al., 2020). To discuss the amount of information a sibling distribution holds, we analyse the relationship between the size of a sibling distribution $\mathcal{D}(w, C_i)$ (word frequency N_i^w) and the rank of a covariance matrix \mathbf{V}_i^w calculated from $\mathcal{D}(w, C_i)$.

Figure 2 shows the relationship between the frequency of randomly sampled 1,000 words and the rank of their covariance matrices. For each word, we construct a covariance matrix from sibling embeddings as in (2). These matrices have $d \times d$ dimensions (BERT base models have $d = 768$ hidden size), and we use their full components (*full(cov)*) for computing their ranks. We see that there is a strong correlation between the frequency and the rank of the covariance matrix, and when the frequency exceeds the dimension size, the rank remains constant at the dimensionality of the contextualised embedding space. This result implies that, upto the dimensionality of the contextualised embedding space, the covariance matrix computed from the sibling distribution $\mathcal{D}(w, C_i)$, retains information about the individual occurrences of a word. Given that contextualised embeddings are often high dimensional (e.g. 768, 1024 etc.) the covariance matrix \mathbf{V}_i^w computed from the sibling

distribution $\mathcal{D}(w, C_i)$ preserves sufficient information about w for semantic variations related to w .

In this analysis, we show that an interesting trend of the word frequency and the rank of covariance matrix. We speculate that this result may be related to the trend of the sense frequency and the length of sense representation reported in the previous study (Zhou and Bollegala, 2022). However, we leave the investigation of this interesting trend to future research.

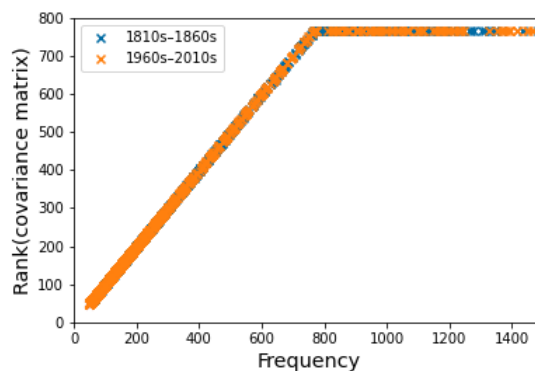


Figure 2: The relationship between the frequency and the rank of the covariance matrix of randomly sampled 1,000 words.

B List of Divergence Measures

We describe the divergence measures as detailed next. For simplicity, we denote two Gaussian distributions $\mathcal{N}(\mu_1^w, \mathbf{V}_1^w)$ and $\mathcal{N}(\mu_2^w, \mathbf{V}_2^w)$ as \mathcal{N}_1^w and \mathcal{N}_2^w , respectively.

Kullback-Liebler

$$\begin{aligned} & \text{KL}(\mathcal{N}_1^w || \mathcal{N}_2^w) \\ &= \frac{1}{2} \left(\text{tr}(\mathbf{V}_2^{w-1} \mathbf{V}_1^w) - d - \log \frac{\det(\mathbf{V}_1^w)}{\det(\mathbf{V}_2^w)} \right. \\ & \quad \left. + (\mu_2^w - \mu_1^w)^\top \mathbf{V}_2^{w-1} (\mu_2^w - \mu_1^w) \right) \end{aligned} \quad (4)$$

Jeffrey's

$$\begin{aligned} & \text{Jeff}(\mathcal{N}_1^w || \mathcal{N}_2^w) \\ &= \frac{1}{2} \text{KL}(\mathcal{N}_1^w || \mathcal{N}_2^w) + \frac{1}{2} \text{KL}(\mathcal{N}_2^w || \mathcal{N}_1^w) \\ &= \frac{1}{4} \left(\text{tr}(\mathbf{V}_2^{w-1} \mathbf{V}_1^w) + \text{tr}(\mathbf{V}_1^{w-1} \mathbf{V}_2^w) - 2d \right. \\ & \quad \left. + (\mu_2^w - \mu_1^w)^\top \mathbf{V}_2^{w-1} (\mu_2^w - \mu_1^w) \right. \\ & \quad \left. + (\mu_1^w - \mu_2^w)^\top \mathbf{V}_1^{w-1} (\mu_1^w - \mu_2^w) \right) \end{aligned} \quad (5)$$

C List of Distance Measures

We describe the distance measures as detailed next. $w(i)$ denotes the i -th value of a word vector w and \bar{w} denotes a subtracted vector from the average of all dimension values.

Bray-Curtis

$$\psi(w_1, w_2) = \frac{\sum_{i \in d} |w_1(i) - w_2(i)|}{\sum_{i \in d} |w_1(i) + w_2(i)|} \quad (6)$$

Canberra

$$\psi(w_1, w_2) = \sum_{i \in d} \frac{|w_1(i) - w_2(i)|}{|w_1(i)| + |w_2(i)|} \quad (7)$$

Chebyshev

$$\psi(w_1, w_2) = \max_i |w_1(i) - w_2(i)| \quad (8)$$

City Block

$$\psi(w_1, w_2) = \sum_{i \in d} |w_1(i) - w_2(i)| \quad (9)$$

Correlation

$$\psi(w_1, w_2) = 1 - \frac{\bar{w}_1 \cdot \bar{w}_2}{\|\bar{w}_1\|_2 \|\bar{w}_2\|_2} \quad (10)$$

Cosine

$$\psi(w_1, w_2) = 1 - \frac{w_1 \cdot w_2}{\|w_1\|_2 \|w_2\|_2} \quad (11)$$

Euclidean

$$\psi(w_1, w_2) = \|w_1 - w_2\|_2 \quad (12)$$

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
6
- A2. Did you discuss any potential risks of your work?
7
- A3. Do the abstract and introduction summarize the paper’s main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

4

- B1. Did you cite the creators of artifacts you used?
4
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We have experimented with the same usage as in the shared task for which the data was provided.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We were unable to find the information for the dataset we used.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4

C Did you run computational experiments?

4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

4

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

4

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

4

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.