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).

As illustrated in Figure 1, we can identify two types of semantic changes associated with words – (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.

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.
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 \( S(w) \) 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 $S(w, C_i)$. 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 $s \in S(w, C_i)$ produced by an MLM $M$ by $f_M(w, s) \in \mathbb{R}^d$, 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 $f_M(w, s)$ as the sibling embeddings of $w$ in context $s$. The number of siblings of $w$ in $C_i$ is denoted by $N_i^w = |S(w, C_i)|$. Moreover, let the set of sibling embeddings of $w$ created from its occurrences in $C_i$ to be $D(w, C_i) = \{f_M(w, s) | s \in S(w, C_i)\}$. 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, $\mu^w_i \in \mathbb{R}^d$ for $w$ by averaging all of its sibling embeddings over the contexts as given by (1).

$$\mu^w_i = \frac{1}{N_i^w} \sum_{s \in S(w, C_i)} f_M(w, s) \quad (1)$$

Although much prior work has used $\mu^w_i$ 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 $\mu^w_i$, we also use the covariance matrix $V^w_i \in \mathbb{R}^{d \times d}$ computed from the sibling embedding distribution of $w$ as defined by (2).

$$V^w_i = \frac{1}{N^w_i (N^w_i - 1)} \sum_{s \in \mathcal{D}(w, C_i)} f_M(w, s) f_M(w, s)^\top$$

We approximate the distribution of sibling embeddings of $w$ using a Gaussian, $\mathcal{N}(\mu^w_i, V^w_i)$ 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 $N^w_i$ and the rank of its covariance matrix $V^w_i$ (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 $\mu^w_i$ and $V^w_i$.

### 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 $\mathcal{N}(\mu^w_1, V^w_1)$ and $\mathcal{N}(\mu^w_2, V^w_2)$. 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

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 Jeffery’s divergence measures using $\mu^w_1, \mu^w_2, V^w_1$ and $V^w_2$ in closed-form formulas (Appendix B).

### Distance measures

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 $\psi(w_1, w_2)$ 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 $w_1$ and $w_2$, we compute the semantic variation score, 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).

$$\text{score}(w) = \frac{1}{N^w_1 N^w_2} \sum_{w_1 \in \mathcal{D}(w, C_1)} \sum_{w_2 \in \mathcal{D}(w, C_2)} \psi(w_1, w_2)$$

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 $\mathcal{N}(\mu^w_1, V^w_1)$ and $\mathcal{N}(\mu^w_2, V^w_2)$. 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 $\mu^w_i$ and $V^w_i$ of the associated sibling distribution.

### 4 Experiments

#### 4.1 Data and Metric

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

---

[It is licensed under a Creative Commons Attribution 4.0 International License.]
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 model\(^3\) (MLM\(_{\text{pre}}\)) and a fine-tuned model (MLM\(_{\text{temp}}\)) from MLM\(_{\text{pre}}\) (Rosin et al., 2022). The base model consists of 12 layers, which we use in two different configurations: (a) we use the last layer (MLM\(_{\text{pre}}|\text{temp, last}\)), and (b) the mean-pool over the last four layers (MLM\(_{\text{pre}}|\text{temp, four}\)), 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 preliminary experiments.

In the prediction of the degree of semantic change for a given word, the set of sibling embeddings for each time period \(D(w, C_1)\) and \(D(w, C_2)\) is acquired from all occurrences in each corpus using the MLM described above, and the distributions across time periods \(\mathcal{N}(\mu_1^w, V_1^w)\) and \(\mathcal{N}(\mu_2^w, V_2^w)\) 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 (\(\text{diag(cov)}\)) in the divergence measures,\(^4\) and both diagonal and full components (\(\text{full(cov)}\)) in the distance measures. Previous studies assume that the covariance matrix is diagonal (\(\text{diag(cov)}\)) (Iwamoto and Yukawa, 2020; Yüksel et al., 2021). This assumption increases computational efficiency compared to \(\text{full(cov)}\), at the expense of losing information on the non-diagonal elements. In our settings, representation of a sibling distribution \(\mathcal{N}(\mu_1^w, V_1^w)\) in \(\text{diag(cov)}\) or \(\text{full(cov)}\) 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\(_{\text{temp}}\)) achieves high performance in all settings. Moreover, for the hidden layers, we have confirmed that our method, by using the last four layers (MLM\(_{\text{pre}}|\text{temp, four}\)), yields even higher correlations than using only the last layer (MLM\(_{\text{pre}}|\text{temp, last}\)).

**Prediction measures.** Our method allows us to try a variety of measures. In the \(\text{diag(cov)}\) 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\(_{\text{temp}}\) (Table 2 and Table 3). Since the cosine distance makes underestimations in MLMs (Zhou

---

\(^3\)https://huggingface.co/bert-base-uncased

\(^4\)In 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.
Table 2: Results of two divergences and seven distance functions under various MLM conditions with the proposed method using \textit{diag}(cov). The best performance in each MLM condition is shown in \textbf{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.

<table>
<thead>
<tr>
<th>Measure</th>
<th>MLM\textsubscript{pre,last}</th>
<th>MLM\textsubscript{pre, four}</th>
<th>MLM\textsubscript{temp,last}</th>
<th>MLM\textsubscript{temp, four}</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL($C_1</td>
<td></td>
<td>C_2$)</td>
<td>0.075</td>
<td>0.130</td>
</tr>
<tr>
<td>KL($C_2</td>
<td></td>
<td>C_1$)</td>
<td>0.100</td>
<td>0.117</td>
</tr>
<tr>
<td>Jeff($C_1</td>
<td></td>
<td>C_2$)</td>
<td>0.090</td>
<td>0.129</td>
</tr>
<tr>
<td>Bray-Curtis</td>
<td>\textbf{0.217}</td>
<td>0.241</td>
<td>0.464</td>
<td>0.480</td>
</tr>
<tr>
<td>Canberra</td>
<td>0.192</td>
<td>0.251</td>
<td>0.455</td>
<td>\textbf{0.517}</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>0.154</td>
<td>0.166</td>
<td>\textbf{0.517}</td>
<td>0.478</td>
</tr>
<tr>
<td>City Block</td>
<td>0.198</td>
<td>0.140</td>
<td>0.461</td>
<td>0.459</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.191</td>
<td>0.266</td>
<td>0.480</td>
<td>0.463</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.190</td>
<td>\textbf{0.270}</td>
<td>0.478</td>
<td>0.480</td>
</tr>
<tr>
<td>Euclidean</td>
<td>0.198</td>
<td>0.249</td>
<td>0.473</td>
<td>0.474</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>MLM\textsubscript{pre,last}</th>
<th>MLM\textsubscript{pre, four}</th>
<th>MLM\textsubscript{temp,last}</th>
<th>MLM\textsubscript{temp, four}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bray-Curtis</td>
<td>\textbf{0.219}</td>
<td>0.263</td>
<td>0.460</td>
<td>0.467</td>
</tr>
<tr>
<td>Canberra</td>
<td>0.195</td>
<td>0.246</td>
<td>0.502</td>
<td>\textbf{0.489}</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>0.145</td>
<td>0.132</td>
<td>\textbf{0.529}</td>
<td>0.451</td>
</tr>
<tr>
<td>City Block</td>
<td>0.192</td>
<td>0.248</td>
<td>0.414</td>
<td>0.452</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.181</td>
<td>\textbf{0.286}</td>
<td>0.481</td>
<td>0.468</td>
</tr>
<tr>
<td>Cosine</td>
<td>0.189</td>
<td>0.272</td>
<td>0.479</td>
<td>0.454</td>
</tr>
<tr>
<td>Euclidean</td>
<td>0.204</td>
<td>0.231</td>
<td>0.454</td>
<td>0.457</td>
</tr>
</tbody>
</table>

**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\textsubscript{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\text{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\text{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\text{temp} model with the average pairwise cosine distance (MLM\text{temp}, APD).

MLM\text{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\text{temp w/ Temp. Att.} (Rosin and Radinsky, 2022): It is the combination of the above two methods (MLM\text{pre, Cosine}), which is considered as the current SoTA model for semantic variation prediction. They 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\text{temp}) outperforms the non-contextualised distribution based methods (Word2Gauss\text{light} and Word2Gauss), and further improvement can be obtained by adding the time-specific attention mechanisms (MLM\text{pre w/ Temp. Att.} and MLM\text{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\text{temp}). Since our method is based on MLM\text{temp}, we compare performance within MLM\text{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\text{temp}, APD) outperforms the cosine distance between average sibling embeddings (MLM\text{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\text{temp, last, diag(cov)}, 0.480 by MLM\text{temp, four, diag(cov)}, and 0.479 by

<table>
<thead>
<tr>
<th>Model</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word2Gauss\text{light}</td>
<td>0.358</td>
</tr>
<tr>
<td>Word2Gauss</td>
<td>0.399</td>
</tr>
<tr>
<td>MLM\text{temp, Cosine}</td>
<td>0.467</td>
</tr>
<tr>
<td>MLM\text{temp, APD}</td>
<td>0.479</td>
</tr>
<tr>
<td>MLM\text{pre w/ Temp. Att.}</td>
<td>0.520</td>
</tr>
<tr>
<td>MLM\text{temp w/ Temp. Att.}</td>
<td><strong>0.548</strong></td>
</tr>
<tr>
<td>Proposed</td>
<td><strong>0.529</strong></td>
</tr>
</tbody>
</table>

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.
MLM\textsubscript{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\textsubscript{temp} w/ Temp. Att. are based on the same model MLM\textsubscript{temp}, their configurations are significantly different. Specifically, MLM\textsubscript{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\textsubscript{temp} and thus does not require additional parameters or training. Although according to Table 4, MLM\textsubscript{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.\footnote{To measure the statistical significance, we use the Fisher transformation (Fisher, 1992).}

### 4.5 Ablation Study

We conduct an ablation study to understand the importance of (i) predicting semantic variation with sibling distributions \(N(\mu^w, V^w)\), and (ii) constructing sibling distributions from the mean \(\mu^w\) and covariance \(V^w\) of sibling embeddings. Based on our best setting Proposed (MLM\textsubscript{temp, last, full(cov)}, Chebyshev), we define two variants: (i) predicting semantic variation score using mean vectors \(\mu^w\) and \(\mu^w\) only as previous studies, and (ii) constructing a sibling distribution with the identity matrix \(N(\mu^w, I)\) instead of the covariance matrix \(V^w\). 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 \((V^w = I\text{ and Proposed})\) eliminate overestimation or underestimation problems in using mean vectors only (w/o \(V^w\)). The variant w/o \(V^w\) 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 vectors 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 \((V^w = I\text{ and Proposed})\) 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*).\footnote{The 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.} Moreover, we find that even with the distribution-based variants, using covariance matrices \(V^w\) computed from sibling embeddings yields even better performance than identity matrices \((V^w = I)\). 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
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


Nina T Ahmad, Lars Borina, and Adam Jatowtb. 2021. Survey of computational approaches to lexical semantic change detection. Computational approaches to semantic change, 6:1.


A Information of Sibling Distribution

In the semantic variation prediction, prior work have applied the mean embeddings $\mu_i^w$ of sibling distribution $D(w, C_i)$ for each word $w$. However, since these methods compress multiple vectors of $D(w, C_i)$ into a single vector $\mu_i^w$, 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 $D(w, C_i)$ (word frequency $N_i^w$) and the rank of a covariance matrix $V_i^w$ calculated from $D(w, C_i)$.

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, up to the dimensionality of the contextualised embedding space, the covariance matrix computed from the sibling distribution $D(w, C_i)$, 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 $V_i^w$ computed from the sibling distribution $D(w, C_i)$ 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 $\mathcal{N}(\mu_1^w, V_1^w)$ and $\mathcal{N}(\mu_2^w, V_2^w)$ as $\mathcal{N}_1^w$ and $\mathcal{N}_2^w$, respectively.

**Kullback-Liebler**

$$\text{KL}(\mathcal{N}_1^w||\mathcal{N}_2^w) = \frac{1}{2} \left( \text{tr}(V_2^{w^{-1}}V_1^{w}) - d - \log \frac{\det(V_1^{w})}{\det(V_2^{w})} \right) + (\mu_2^w - \mu_1^w)^\top V_2^{w^{-1}}(\mu_2^w - \mu_1^w)$$ (4)

**Jeffrey’s**

$$\text{Jeff}(\mathcal{N}_1^w||\mathcal{N}_2^w) = \frac{1}{2} \text{KL}(\mathcal{N}_1^w||\mathcal{N}_2^w) + \frac{1}{2} \text{KL}(\mathcal{N}_2^w||\mathcal{N}_1^w) = \frac{1}{4} \left( \text{tr}(V_2^{w^{-1}}V_1^{w}) + \text{tr}(V_1^{w^{-1}}V_2^{w}) - 2d + (\mu_2^w - \mu_1^w)^\top V_2^{w^{-1}}(\mu_2^w - \mu_1^w) + (\mu_1^w - \mu_2^w)^\top V_1^{w^{-1}}(\mu_1^w - \mu_2^w) \right)$$ (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?

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?

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.)?

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

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.?

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)?

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?

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

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