A Semantic Distance Metric Learning approach for Lexical Semantic Change Detection

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

Danushka Bollegala
University of Liverpool
danushka@liverpool.ac.uk

Abstract

Detecting temporal semantic changes of words is an important task for various NLP applications that must make time-sensitive predictions. Lexical semantic change detection (SCD) task involves predicting whether a given target word, w, changes its meaning between two different text corpora, C_1 and C_2. For this purpose, we propose a supervised two-staged SCD method that uses existing Word-in-Context (WiC) datasets. In the first stage, for a target word w, we learn two *sense-aware* encoders that represent the meaning of w in a given sentence selected from a corpus. Next, in the second stage, we learn a *sense-aware* distance metric that compares the semantic representations of a target word across all of its occurrences in C_1 and C_2. Experimental results on multiple benchmark datasets for SCD show that our proposed method achieves strong performance in multiple languages. Additionally, our method achieves significant improvements on WiC benchmarks compared to a sense-aware encoder with conventional distance functions.[1]

1 Introduction

The notion of word meaning is a dynamic one, and evolves over time as noted by Tahmasebi et al. (2021). For example, the meaning of the word *cell* has changed over time to include *cell phone* to its previous meanings of *prison* and *related to biology*. Detection of such semantic changes of words over time remains a challenging, yet an important task for lexicography, sociology, and information retrieval (Traugott and Dasher, 2001; Cook and Stevenson, 2010; Michel et al., 2011; Kutuzov et al., 2018). For example, in E-commerce, a user might use the same keyword, such as *scarf*, to search for different types of products based on seasonal variations, such as *silk scarves in spring* versus *woollen scarves in winter*. Recently, a decline in the performance of pretrained Large Language Models (LLMs) over time has been observed, and attributed to their training on static snapshots (Loureiro et al., 2022; Lazaridou et al., 2021). Recognising the words with changed meanings enables efficient fine-tuning of LLMs to incorporate only those with specific semantic shifts (yu Su et al., 2022).

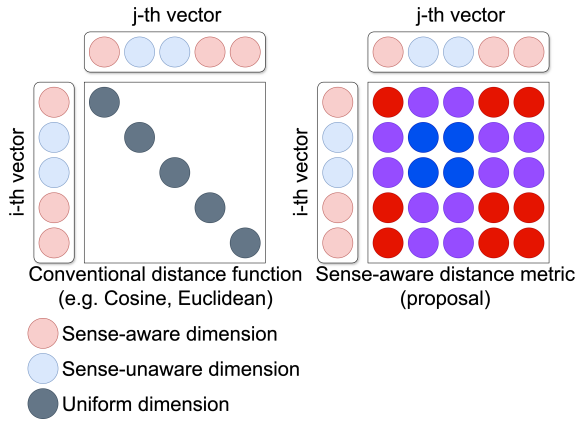

Figure 1: An overview of our method. Conventional distance functions such as Cosine or Euclidean consider the information of each dimension uniformly (left), but our method considers sense-aware information and cross-dimensional correlations (right).

Detecting whether a word has its meaning changed between two given corpora, sampled at different points in time, requires overcoming two important challenges, which we name as the *representational* and *measurement* challenges.

Representational Challenge: A word can take different meanings in different contexts even within the same corpus. Therefore, creating a representation for the meaning of a word across an entire corpus is a challenging task compared to that in a single sentence or a document. Prior work has averaged static (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016; Yao et al., 2018; Du-

[1]Source code is available at https://github.com/LivNLP/svp-sdml .

Findings of the Association for Computational Linguistics: ACL 2024, pages 7570–7584

bossarsky 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) word embeddings for this purpose, which is suboptimal because averaging conflates multiple meanings of a word into a single vector.

Measurement Challenge: We require a distance metric that can accurately measure the semantic change of a target word between two given corpora, using the learnt representations of the target word from each corpus. This challenge has been addressed in prior work using parameter-free distance functions due to the lack of labelled training data to provide any supervision for SCD (Kutuzov and Giulianelli, 2020; Card, 2023; Cassotti et al., 2023; Aida and Bollegala, 2023b,a; Tang et al., 2023).

To address those challenges, we propose Semantic Distance Metric Learning (SDML), a supervised two-staged SCD method. To solve the representational challenge, we learn a sense-aware[2] encoder from sense-level supervision that represents the meaning of a target word in a sentence. Word sense information has shown to be useful for SCD (Rachinskiy and Arefyev, 2021; Arefyev et al., 2021; Cassotti et al., 2023; Tang et al., 2023) and outperform word representations that encode only time (Rosin et al., 2022; Rosin and Radinsky, 2022). Following Cassotti et al. (2023), we use WiC (Pilehvar and Camacho-Collados, 2019) data, annotated for word sense discrimination for training the sense-aware encoder.

To solve the measurement challenge, we propose a method to learn a sense-aware distance metric to compare two sense-aware embeddings for a target word in sentences selected from the two corpora. Specifically, the distance metric is trained such that it returns a smaller value between two sentences where the target word takes the same meaning compared to that between two sentences which express different meanings of the target word. We learn a distance metric that satisfies this criteria, using the weak-supervision provided by existing WiC datasets.

[2]In this paper, the term 'sense' refers not to strictly defined senses as found in dictionaries or WordNet (Fellbaum and Miller, 1998), but rather to the nuances of word meanings derived from the distributional hypothesis (Harris, 1954).

Experimental results show that SDML achieves exceptional performance in the SCD task. Remarkably, SDML obtains performance improvements of 2-5% over the current strong baselines. While our focus is on the SCD task, we also explore the applicability of our SDML to the WiC task. In WiC benchmarks, SDML significantly enhances performance in multiple languages. These results demonstrate the effectiveness of learning both a sense-aware encoder and a sense-aware distance metric.

2 Related Work

The diachronic semantic changes of words have been extensively investigated by linguists to explore how meanings evolve over time (Traugott and Dasher, 2001). In recent years, the advent of diachronic corpora and advancements in representing word meanings have paved the way for automated SCD in NLP (Tahmasebi et al., 2021). To detect semantically changed words in time-separated corpora, unsupervised SCD methods compare word embeddings trained on the target corpora. Many methods exist that align vector spaces over time spanned by static word embeddings, 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). Likewise, methods for comparing sets of target word representations computed as contextualised word embeddings have also been proposed. Such methods include comparing the average (Martinc et al., 2020; Beck, 2020; Kutuzov and Giulianelli, 2020; Laicher et al., 2021; Giulianelli et al., 2022; Rosin et al., 2022; Rosin and Radinsky, 2022) or each pair of embeddings (Kutuzov and Giulianelli, 2020; Laicher et al., 2021). Additionally, Aida and Bollegala (2023b) and Nagata et al. (2023) have proposed methods that consider the variance in the sets of embeddings. These unsupervised SCD methods have been evaluated using unsupervised SCD tasks (Del Tredici et al., 2019; Schlechtweg et al., 2020; Kutuzov and Pivovarova, 2021).

Due to the lack of manually-labelled data for lexical semantic change, prior work on SCD have been limited to unsupervised approaches. A notable exception is XL-LEXEME (Cassotti et al., 2023), a supervised SCD method that uses sense-level supervision. They focused on the WiC task (Pilehvar and Camacho-Collados, 2019) de-

signed to detect semantic differences of a target word in a given pair of sentences. XL-LEXEME fine-tunes a pretrained multilingual Masked Language Model (MLM) across multilingual WiC datasets, and achieves strong performance in some SCD tasks. Recent work show SCD is a challenging task for LLMs such as GPT-3.5 (Sorensen et al., 2022; Periti et al., 2024), reporting low performance even with carefully designed prompts.

3 Semantic Distance Metric Learning

Given a target word w and two corpora C_1 and C_2 sampled at distinct time points, the goal in SCD is to predict a score for w that indicates the degree of semantic change undergone by w between C_1 and C_2. For this purpose, we propose, SDML, which overcomes the two challenges introduced in § 1. Specifically, to address the representational challenge, we first learn a sense-aware encoder in § 3.1. Next, to overcome the measurement challenge, we learn a sense-aware distance metric between two sense-aware representations of a target word computed from C_1 and C_2, as described in § 3.2.

3.1 Learning Sense-Aware Encoder

Following prior work on SCD that show contextualised word embeddings to outperform the static word embeddings (Aida and Bollegala, 2023b; Rosin and Radinsky, 2022), we represent the meaning of a word w in a sentence s by the d-dimensional token embedding,[3] $\boldsymbol{f}(w, s; \theta)(\in \mathbb{R}^d)$, obtained from a pretrained MLM parametrised by θ. Although contextualised word embeddings capture the contextual information of a word in a sentence (Zhou and Bollegala, 2021), explicitly encoding word sense information has been shown to improve performance in SCD tasks (Tang et al., 2023).

For this purpose, we follow Cassotti et al. (2023) and fine-tune the MLM on WiC. WiC contains tuples (w, s_1, s_2, y), where the label $y = 1$ if w has the same meaning in both sentences s_1 and s_2, and $y = 0$ otherwise. We use a Siamese bi-encoder approach (Reimers and Gurevych, 2019), which uses two encoders to produce respective embeddings for pairs of sentences. This architecture enables us to capture the semantic relationship between the pairs of sentences. In this paper, we use the same MLM encoder to obtain two representations, $\boldsymbol{w}_1 = \boldsymbol{f}(w, s_1; \theta)$ and $\boldsymbol{w}_2 = \boldsymbol{f}(w, s_2; \theta)$ for the meaning of w in s_1 and s_2. We then update θ such that the contrastive loss, ℓ_c (Hadsell et al., 2006) given by (1) is minimised.

$$\ell_c(\theta) = \frac{1}{2}\left(y\delta^2 + (1-y)\max(0, m-\delta)^2\right) \quad (1)$$

Here, we set the margin $m = 0.5$ and $\delta = 1 - (\boldsymbol{w}_1{}^\top \boldsymbol{w}_2)/\left|\left|\boldsymbol{w}_1\right|\right|\left|\left|\boldsymbol{w}_2\right|\right|$ is the cosine distance. Note that the same encoder is used to compute both $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ here. We use AdamW (Loshchilov and Hutter, 2019) as the optimiser to minimise the contrastive loss in (1) with the initial learning rate set to 10^{-5} and the weight decay coefficient set to 0.01.

3.2 Learning Sense-Aware Distance Metrics

Armed with the sense-aware encoder trained in § 3.1, we are now ready to learn a sense-aware distance metric $h(\boldsymbol{w}_1, \boldsymbol{w}_2; \mathbf{A})$ given by (2) that measures the semantic distance between two sense-aware embeddings $\boldsymbol{w}_1 = \boldsymbol{f}(w, s_1; \theta)$ and $\boldsymbol{w}_2 = \boldsymbol{f}(w, s_2; \theta)$, for w in two sentences s_1 and s_2, respectively.

$$h(\boldsymbol{w}_1, \boldsymbol{w}_2; \mathbf{A}) = (\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \mathbf{A}(\boldsymbol{w}_1 - \boldsymbol{w}_2) \quad (2)$$

Here, $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a Mahalanobis matrix (assumed to be positive definite) that defines a (squared) distance metric, Mahalanobis distance. Unlike conventional distance functions such as Euclidean distance, Mahalanobis distance weights each dimension according to its variance and accounts for cross-dimensional correlations.

We use the WiC training data instances to learn $\mathbf{A}$ in (2). Specifically, we consider two types of constraints that must be satisfied by h as follows. For instances where $y = 1$, the distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is required to be greater than a distance lower bound ℓ such that, $h(\boldsymbol{w}_1, \boldsymbol{w}_2) \geq \ell$. Likewise, for instances where $y = 0$, the distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is required to be lower than a distance upper bound u such that, $h(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq u$. There exists a bijection (up to a scaling function) between the set of Mahalanobis distances and the set of equal mean (denoted by $\boldsymbol{\mu}$) multivariate Gaussian distributions given by $p(\boldsymbol{w}; \mathbf{A}) = \frac{1}{Z}\exp\left(-\frac{1}{2}h(\boldsymbol{w}, \boldsymbol{\mu}; \mathbf{A})\right)$, where Z is a normalising constant and $\mathbf{A}^{-1}$ is the covariance of the distribution. We can use this bijection to measure the distance between two Mahalanobis distance functions parametrised by positive definite

[3] When a word has been tokenised into multiple subtokens, we compute the average of the subtoken embeddings to create the token embedding

matrices $\mathbf{A}_0$ and $\mathbf{A}$ using the relative entropy (KL-divergence) between the corresponding multivariate Gaussians, given by (3).

$$\mathrm{KL}(p(\boldsymbol{w};\mathbf{A}_0)||p(\boldsymbol{w};\mathbf{A})) = \int p(\boldsymbol{w};\mathbf{A}_0)\log\frac{p(\boldsymbol{w};\mathbf{A}_0)}{p(\boldsymbol{w};\mathbf{A})}d\boldsymbol{w} \tag{3}$$

Relative entropy is a non-negative and convex function in $\mathbf{A}$. The overall optimisation problem is then given by (4).

$$\begin{aligned} \min_{\mathbf{A}} \quad & \mathrm{KL}(p(\boldsymbol{w};\mathbf{A}_0)||p(\boldsymbol{w};\mathbf{A})) \\ \text{subject to} \quad & h(\boldsymbol{w}_1,\boldsymbol{w}_2) \leq u \quad y=1, \\ & h(\boldsymbol{w}_1,\boldsymbol{w}_2) \geq \ell \quad y=0. \end{aligned} \tag{4}$$

The optimisation problem given in (4) can be expressed as a particular type of Bregman divergence, which can be efficiently solved using the Bregman's method (Censor and Zenios, 1997). We use Information-Theoretic Metric Learning (ITML) (Davis et al., 2007) for learning a Mahalanobis matrix $\mathbf{A}$ that satisfies those requirements. Further details of ITML are described in Appendix A.

3.3 Measuring Temporal Semantic Change

Using the sense-aware distance metric h and the sense-aware encoder f learnt as described in the previous sections, we can now compute the semantic change score, $\alpha(w, C_1, C_2)$, of w between C_1 and C_2 as the average pairwise distance computed over $\mathcal{S}_1(w)$ and $\mathcal{S}_2(w)$, given by (5).

$$\frac{1}{n_1 n_2}\sum_{\substack{s_1\in\mathcal{S}_1(w)\\ s_2\in\mathcal{S}_2(w)}} h(\boldsymbol{f}(w,s_1;\theta),\boldsymbol{f}(w,s_2;\theta);\mathbf{A}) \tag{5}$$

Here, we denote the set of sentences where w occurs in C_i to be $\mathcal{S}_i(w)$ for $i = 1, 2$, and the corresponding number of sentences in each set to be $n_i = |\mathcal{S}_i(w)|$. Unlike much prior work in SCD, which first computes a single vector (often by averaging) for a target word w from a corpus, thereby conflating its different meanings, (5) computes representations per each occurrence of w and compares the average over all distances. Although the total number of summations, $n_1 n_2$, in (5) could become large for frequent w, semantic change scores can be efficiently computed by pre-computing and indexing the sense-aware embeddings for all occurrences of w in each corpus. Moreover, the summation in (5) can be computed in parallel over different batches of sentences to obtain a highly efficient Map-Reduce version (Dean and Ghemawat, 2004).

Dataset	Language	#Train	#Dev	#Test
XL-WiC	German	48k	8.9k	1.1k
	French	39k	8.6k	22k
	Italian	1.1k	0.2k	0.6k
MCL-WiC	English	4.0k	0.5k	0.5k
AM^{2}iCo	German	50k	0.5k	1.0k
	Russian	28k	0.5k	1.0k
	Japanese	16k	0.5k	1.0k
	Chinese	13k	0.5k	1.0k
	Arabic	9.6k	0.5k	1.0k
	Korean	7.0k	0.5k	1.0k
	Finnish	6.3k	0.5k	1.0k
	Turkish	3.9k	0.5k	1.0k
	Indonesian	1.6k	0.5k	1.0k
	Basque	1.0k	0.5k	1.0k

Table 1: Statistics of the WiC datasets. #Train, #Dev, and #Test shows the number of instances.

4 Experiments

4.1 Setting

To learn the sense-aware encoder and the distance metric described in § 3, we use WiC datasets covering multiple languages as shown in Table 1: XL-WiC (Raganato et al., 2020), MCL-WiC (Martelli et al., 2021), and AM^{2}iCo (Liu et al., 2021).[4]

We use the sense-aware encoder released by Cassotti et al. (2023) as f,[5] which is based on XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2020) for the remainder of the experiments with SDML reported in this paper. We use the `metric_learn`[6] package to train a sense-aware distance metric with ITML. The slack parameter γ in ITML weights the margin violations, and is searched from the 11 values in $\{10^{-5}, 10^{-4}, ..., 10^{0}, ..., 10^{4}, 10^{5}\}$, according to the best performance measured on the development data in each WiC dataset.

After that, we evaluate the performance on SCD tasks. In this paper, we use the two benchmark datasets – SemEval-2020 Task 1 (Schlechtweg et al., 2020) (covering English (En), German (De), Swedish (Sv) and Latin (La)) and RuShiftEval (Kutuzov and Pivovarova, 2021) (covering Russian (Ru)), which have also been used in prior work on SCD (Kutuzov et al., 2021; Giulianelli

[4] XL-WiC and MCL-WiC are licensed under a Creative Commons Attribution-NonCommercial 4.0 License, and AM^{2}iCo is licensed under a Creative Commons Attribution 4.0 International Public license.

[5] This model is available at https://huggingface.co/pierluigic/xl-lexeme

[6] https://github.com/scikit-learn-contrib/metric-learn

Dataset	Language	Time Period	#Sentences
SemEval	English	1810–1860	254k
		1960–2010	354k
	German	1800–1899	2.6M
		1946–1990	3.5M
	Swedish	1790–1830	3.4M
		1895–1903	5.2M
	Latin	B.C. 200–0	96k
		0–2000	463k
RuShiftEval	Russian	1700–1916	3.3k
		1918–1990	3.3k
		1992–2016	3.3k

Table 2: Statistics of the SCD datasets. In the RuShiftEval, we used annotated pairwise sentences for prediction as in Cassotti et al. (2023), and evaluation is conducted in three subsets: pre-Soviet (1700–1916) vs. Soviet (1918–1990), Soviet (1918–1990) vs. post-Soviet (1992–2016), and pre-Soviet (1700–1916) vs. post-Soviet (1992–2016), respectively referred to as RuShiftEval1 (Ru_1), RuShiftEval2 (Ru_2), and RuShiftEval3 (Ru_3). Full data statistics are shown in Appendix B.

et al., 2022; Cassotti et al., 2023).[7] Statistics of those datasets are summarised in Table 2.

For English, German and Russian, there exist WiC training data splits that we can use to train SDML separately for each of those languages. However, no such training data are available for Latin and Swedish languages. Therefore, when evaluating SDML for Latin, we train it on the WiC training data available for Italian and French, which are in the same Romance language family. Likewise, when evaluating SDML for Swedish, we train it on the WiC training data available for German, which is in the same Germanic language family. Creating WiC datasets for languages that do not have such resources is beyond the scope of this paper and is deferred to future work.

Before presenting the results of the SCD task, we first look at the performance of our SDML on the WiC task for languages corresponding to the SCD benchmarks. As a baseline, we use a sense-aware encoder fine-tuned to distinguish the different meanings of a target word by optimising (1). For SDML, the sense-aware distance metric is learned via ITML using WiC datasets, classification boundaries are also obtained. Therefore, we can use these boundaries for the prediction of the WiC tasks.

[7]SemEval-2020 Task 1 and RuShiftEval are licensed under a Creative Commons Attribution 4.0 International License and a GNU General Public License version 3.0, respectively.

Table 3 shows that the combination of the sense-aware encoder and the sense-aware distance metric (our SDML) constantly outperforms the baseline. Interestingly, SDML achieves significant improvements at the 95% confidence interval computed using Bernoulli trials in all languages. These results support our hypothesis: even in the task of detecting semantic differences at the same time period, it is better to use the combination of the sense-aware encoder and the sense-aware distance metric than the sense-aware encoder only. More comprehensive results can be found in § 4.3.

4.2 Evaluating Semantic Changes of Words

We compare the performance of SDML against prior work on several SCD benchmarks. In this evaluation, given a set of target words, an SCD method under evaluation is required to predict scores that indicate the degree of semantic changes undergone by each word in the set. The Spearman's rank correlation coefficient r ($\in [-1, 1]$) between those predicted semantic change scores and that assigned by the human annotators in each benchmark dataset (i.e. *gold* ratings) is computed as the evaluation metric. An SCD method that reports a high Spearman's r value indicates better agreement with human ratings, and is considered to be desirable for detecting the semantic changes of words over time.

We compare the proposed methods against strong baselines as described next.

Baselines: Kutuzov and Giulianelli (2020) and Laicher et al. (2021) showed that the average pairwise cosine distance of the sets of contextualised embeddings perform well (**APD**). Aida and Bollegala (2023a) proposed swapping-based SCD, conducting context-swapping across target corpora to obtain more reliable scores (**SSCD**). Card (2023) proposed a token replacement-based JS divergence metric, to mitigate the influence of token frequency by replacing tokens with neighbouring words (**ScaledJSD**). Rosin and Radinsky (2022) proposed temporal attention, additional time-aware attention mechanism to MLMs. SCD scores are calculated by the average cosine distance (**w/ TA + CD**). Giulianelli et al. (2022) introduced an ensemble method combining the average cosine distance from MLMs with the cosine distance between morpho-syntactic features labelled from universal dependencies (**CD & UD + CD**).

Models	MCL-WiC	XL-WiC			AM^{2}iCo	
	En	De	Fr	It	De	Ru
Baseline: Sense-aware Sentence Encoder	78.0	78.3	73.2	67.1	78.1	78.2
+ Sense-aware Distance Metric	**90.3**††	**84.9**††	**78.7**††	**75.3**††	**85.0**††	**87.6**††

Table 3: Accuracy of the WiC test sets for languages relevant to the SCD benchmarks. †† indicates significance at the 95% confidence interval.

Models	FT	SemEval				RuShiftEval		
		En	De	Sv	La	Ru_1	Ru_2	Ru_3
Baselines: MLM								
+ APD (Laicher et al., 2021)		0.571	0.407	0.554	-	-	-	-
+ SSCD (Aida and Bollegala, 2023a)		0.383	0.597	0.234	0.433	-	-	-
+ APD (Kutuzov and Giulianelli, 2020)	✓	0.605	0.560	0.569	0.113	-	-	-
+ ScaledJSD (Card, 2023)	✓	0.547	0.563	0.310	0.533	-	-	-
w/ TA + CD (Rosin and Radinsky, 2022)	✓	0.520	0.763	-	0.565	-	-	-
+ CD & UD + CD (Giulianelli et al., 2022)	✓	0.451	0.354	0.356	**0.572**	0.117	0.269	0.326
Sense-Aware Methods								
DeepMistake (Arefyev et al., 2021)	✓	-	-	-	-	0.798	0.773	0.803
GlossReader (Rachinskiy and Arefyev, 2021)	✓	-	-	-	-	0.781	0.803	0.822
XL-LEXEME (Cassotti et al., 2023)		0.757	0.877	**0.754**	0.056	0.775	0.822	0.809
XL-LEXEME (Cassotti et al., 2023)	✓	-	-	-	-	0.799	**0.833**	0.842
SDML (ours)		**0.774**	**0.902**	0.656	0.124	**0.805**	0.811	**0.846**

Table 4: Spearman's rank correlation on SCD tasks compared against strong baselines. FT indicates whether fine-tuning on target time separated corpora was conducted. The absolute correlations for previous methods are taken from the respective papers as reported. - indicates that the corresponding benchmark was not evaluated in the referenced paper.

Sense-Aware Methods: Here, we introduce sense-aware methods that use sense-level supervision. **DeepMistake** (Arefyev et al., 2021) and **GlossReader** (Rachinskiy and Arefyev, 2021) both leverage data from the WiC and Word Sense Disambiguation tasks to fine-tune MLMs. Subsequently, both methods make predictions using linear regression trained on the provided training data. **XL-LEXEME** is fine-tuned on WiC data across multiple languages. In our method, SCD is performed using the average pairwise distance. For SDML, we predict the semantic change scores for the target words using (5).

Results are shown in Table 4.[8] Our method performs comparable or superior to previous sense-aware methods using a sense-aware encoder only. Moreover, SDML accomplishes performance improvements of 2-5% over the sense-aware methods[9], provided that training data in the corresponding languages is available. In SemEval Sv and La, no training data for the corresponding languages exist, but our method performed equally well for SemEval Sv and showed a drastic performance improvement for SemEval La.

However, the performance in Latin is lower than pretrained MLMs in Table 4. We attribute these results to the composition of the data used during MLM pretraining and SDML training, respectively. Firstly, while Latin is included in the pretraining of XLM-RoBERTa, the dataset for pretraining is only a fraction (1/10 to 1/100) of the size of the datasets for the other languages evaluated in the SCD benchmarks. Secondly, Latin is also absent from the WiC dataset used to train our SDML. Therefore, we believe that the poor information on Latin, which was scarce even during the pretraining phase, was further diluted by the training of the SDML, resulting in a low performance. In the future, as the WiC datasets are expanded and include more languages, it is expected that the performance of models for a wider range of languages will improve.

4.3 Applying to WiC Prediction

Our main focus in this paper has been Lexical Semantic Change Detection – detecting whether a target word, w, has its meaning changed from

[8] In SemEval De, Sv, and La, we have several models due to multiple WiC datasets in the corresponding/related languages, but we report the highest performance.

[9] In SCD tasks, there is no statistical significance across ALL methods due to the lack of target words when we use the Fisher transformation.

Models	MCL-WiC	XL-WiC all			XL-WiC IV			XL-WiC OOV		
	En	De	Fr	It	De	Fr	It	De	Fr	It
Baselines: MLM										
$\text{mBERT}_{\text{base}}$	84.0	81.6	73.7	72.0	81.9	72.9	73.2	70.1	71.2	68.5
$\text{XLM-RoBERTa}_{\text{base}}$	86.6	80.8	73.1	68.6	81.2	71.9	70.7	71.3	71.1	62.4
$\text{XLM-RoBERTa}_{\text{large}}$	-	84.0	76.2	72.3	82.2	75.6	75.1	72.5	73.9	65.2
Lang. BERT	-	82.9	78.1	72.6	83.2	77.6	73.9	**76.6**	**78.0**	69.1
Sense-Aware Methods										
XL-LEXEME	78.0	78.3	73.2	67.1	78.4	73.6	71.2	65.2	65.7	57.9
SDML (ours)	**90.3**†	**84.9**†	**78.7**††	**75.3**	**85.1**†	**78.5**††	**77.8**	**76.6**	75.3	**70.2**

Table 5: Accuracy reported by different methods on the MCL-WiC and XL-WiC datasets. XL-WiC contains additional two test sets; in-vocabulary (IV) and out-of-vocabulary (OOV) test sets. Lang. BERT means language-specific BERT models. † or †† indicate significance at the 90% or 95% confidence interval, respectively.

Models	De	Ru	Ja	Zh	Ar	Ko	Fi	Tr	Id	Eu
Baselines: MLM										
$\text{mBERT}_{\text{base}}$	80.4	82.1	78.2	75.2	73.3	75.8	81.2	80.6	78.4	75.9
$\text{XLM-RoBERTa}_{\text{base}}$	79.4	80.9	79.4	76.1	73.6	76.0	81.2	80.5	77.9	74.2
Sense-Aware Methods										
XL-LEXEME	78.1	78.2	77.1	75.2	75.4	75.5	78.0	78.7	75.5	72.7
SDML (ours)	**85.0**†	**87.6**††	**82.9**	**81.7**††	**81.8**††	**82.9**††	**87.7**††	**84.4**	**83.6**††	**80.8**†

Table 6: Accuracy of AM^{2}iCo dataset. † or †† indicate significance at the 90% or 95% confidence interval, respectively.

one corpus, C_1, to another, C_2. However, we use WiC datasets for training the sense-aware encoder (described in § 3.1) as well as the sense-aware distance metric (described in § 3.2) because there does not exist sufficiently large manually annotated datasets for training temporal SCD methods. Although none of the WiC datasets we used are sampled from temporally distinct corpora, they provide a convenient alternative for training models that discriminate the meaning of a word in two given sentences. Therefore, by evaluating our proposed SDML on benchmark datasets for WiC, we will be able to sanity check whether SDML can indeed detect words that have same/different meanings in two given sentences, even if the two sentences might not be sampled from temporally distinct corpora. In this WiC task, a model is required to predict whether a target word takes the same meaning in two given sentences. This is modelled as a binary classification task and classification accuracy is used as the evaluation metric. A random prediction baseline would report an accuracy of 50% on WiC test datasets, which are balanced.

Baselines: As baselines for comparison, use report the accuracy of binary classifiers that have been trained using different MLMs for the WiC task as follows. Given an instance (w, s_1, s_2, y) from a WiC training dataset, the token embeddings $\boldsymbol{f}(w, s_1)$ and $\boldsymbol{f}(w, s_2)$ are concatenated to represent the meaning of w in s_1 and s_2. Next, a binary logistic regression classifier is trained on the training data from a WiC dataset. Note that the parameters of the MLMs are not updated during this process. For these baselines, we use the multilingual MLMs such as $\text{mBERT}_{\text{base}}$ and $\text{XLM-RoBERTa}_{\text{base/large}}$ as well as target language-specific BERT models (Lang. BERT) to train the baselines in XL-WiC such as BERT-base-german-cased[10] for German, CamemBERT-large[11] for French, and BERT-base-italian-xxl-cased[12] for Italian. Rather than re-implementing or re-running these methods, we use the results presented in the corresponding benchmarks.

Sense-Aware Methods: Unlike the MLMs used in the above-mentioned baselines, XL-LEXEME uses a sense-aware encoder fine-tuned to discriminate the different meanings of a target word. XL-LEXEME is already fine-tuned using WiC datasets and does not require a binary classifier to

[10] https://huggingface.co/dbmdz/bert-base-german-cased

[11] https://huggingface.co/almanach/camembert-large

[12] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

be trained, and predicts cosine distances less than the margin m (set to 0.5) to be instances where the target word takes the same meaning in both sentences. For SDML, we use the sense-aware distance metric learned via ITML to predict whether a target word expresses the same meaning in the two given sentences in WiC datasets. Since ITML also learns classification boundaries (i.e. the upper and lower bounds denoted by respectively u and l in Algorithm 1), which can be used to make binary predictions. Note that however, we do not have two different corpora in this WiC-based evaluation, thus not requiring to average the distances as done in (5), and instead can perform a single point estimate considering the two sentences in a single WiC test instance.

Table 5 shows the results for MCL-WiC and XL-WiC datasets. XL-WiC has in-vocab (IV) test sets and out-of-vocab (OOV) test sets extracted from the test set, in addition to the vanilla test set. The results show that our SDML consistently outperforms all of the baselines. It shows 90% accuracy in English, which is a significant improvement compared to XL-LEXEME, using the sense-aware distance metric. In German and French, SDML outperforms XLM-RoBERTa$_{\text{large}}$, which is the same model as XL-LEXEME, as well as language-specific language models (Lang. BERT). Interestingly, even in Italian, where there are only a few thousand training data instances (Table 1), the performance was significantly improved from XL-LEXEME. Moreover, statistical tests reveal that our SDML achieves significant improvements over all other methods on multiple test sets, with statistical significance observed at the 90% or 95% for binomial proportion confidence intervals.

Similarly, in the AM^{2}iCo dataset, Table 6 also demonstrates the performance improvements achieved by SDML. In particular, according to Table 6, SDML achieves the best performance on all test sets. Statistical tests demonstrate the significant performance enhancements of our SDML compared to the other methods on most test sets, with significance observed at the 90% or 95% for binomial proportional confidence intervals. While regular multilingual MLMs and XL-LEXEME do not perform well for low-resource language (Eu) or languages not similar to English (Ja, Zh, Ar, and Ko), sense-aware distance metrics can dramatically improve performance for even just a few thousand instances.

5 Ablation Study

Motivated by the superior performance reported in § 4.2 by SDML, we conduct an ablation study to understand the importance of (i) using sense-aware distance metric, and (ii) considering cross-dimensional correlations by Mahalanobis matrix.

Firstly, we conduct a quantitative comparison between our SDML (sense-aware sentence encoder and sense-aware distance metric using *full* components of Mahalanobis matrix, SDML$_{\text{full}}$) and two variants; (i) sense-aware sentence encoder only (Baseline), and (ii) sense-aware sentence encoder and sense-aware distance metric using *diagonal* components of Mahalanobis matrix (SDML$_{diag}$). Results on SCD tasks are shown in Table 7. While our method achieves comparable or superior performance to the baseline even with the diagonal components of the Mahalanobis matrix (SDML$_{diag}$), using the full Mahalanobis matrix (SDML$_{\text{full}}$) yields further improvements. These results indicate the importance of considering inter-dimensional information by the full components of the sense-aware Mahalanobis matrix.

Secondly, we perform a qualitative analysis using gold labels. Following Aida and Bollegala (2023a,b), we pick up the (i) top-8 semantically changed words, and (ii) top-8 semantically stable words in the SemEval En. Additionally, due to the sense-aware method, we also evaluated the relationship between the SCD performance and two factors; polysemy and frequency as described in Hamilton et al. (2016). We count the number of senses (synsets defined in WordNet) and frequency of the evaluation set of words in given corpora. From Table 8, we can see that our method SDML$_{\text{full}}$ slightly improves prediction. Moreover, polysemous words tend to change their meanings (the law of innovation), but frequent words, contrary to the law of conformity, show no correlation with semantic change, as described in (Hamilton et al., 2016). The same trends are in Table 9. From this table, SDML$_{diag}$ has a higher/lower correlation with frequency/polysemy than the baseline, which degrades the performance in SCD. However, in SDML$_{\text{full}}$, while a correlation with frequency/polysemy is slightly weaker than the baseline, they contribute to the performance improvement in SCD. We conclude that reliev-

[13] http://wordnetweb.princeton.edu/perl/webwn

Models	SemEval			RuShiftEval			
	En	De	Sv	La	Ru_1	Ru_2	Ru_3
Baseline: Sense-aware Sentence Encoder	0.757	0.877	**0.754**	0.056	0.775	**0.822**	0.809
+ Sense-aware Distance Metric (diagonal)	0.750	**0.902**	0.642	0.083	0.804	0.808	**0.846**
+ Sense-aware Distance Metric (full)	**0.774**	**0.902**	0.656	**0.124**	**0.805**	0.811	**0.846**

Table 7: Spearman's rank correlation on SCD tasks.

Word	Gold rank	Δ	WordNet #Synsets	Frequency C_1	Frequency C_2	Baseline rank	$SDML_{diag}$ rank	$SDML_{full}$ rank
plane	1	✓	5	278	792	1	1	1
tip	2	✓	9	119	241	3	6	4
prop	3	✓	3	121	147	9	7	7
graft	4	✓	3	119	109	7	2	6
record	5	✓	8	420	1188	5	4	2
stab	7	✓	3	92	117	4	8	9
bit	9	✓	11	296	622	13	14	13
head	10	✓	33	3599	4127	14	18	14
multitude	30	✗	3	475	131	27	26	24
savage	31	✗	2	504	133	15	13	17
contemplation	32	✗	2	240	111	28	27	28
tree	33	✗	3	2322	1596	32	32	35
relationship	34	✗	4	130	841	24	23	27
fiction	35	✗	2	202	326	35	33	33
chairman	36	✗	1	147	683	36	36	34
risk	37	✗	4	286	643	37	37	37
Spearman	1.000		0.420	−0.153	−0.047	0.757	0.750	**0.774**

Table 8: Ablation study on the words categorised by the existence of semantic change: highlighting the top-8 semantically changed words with significant semantic change ($\Delta = ✓$) and bottom-8 stable words with minimal semantic change ($\Delta = ✗$) on SemEval En. Baseline is a sense-aware sentence encoder only. #Synsets shows the number of synsets in WordNet.[13]

Models	Gold	WordNet #Synsets	Frequency C_1	Frequency C_2
Baseline	0.757	0.427	-0.182	-0.062
$SDML_{diag}$	0.750	0.355	-0.205	-0.121
$SDML_{full}$	**0.774**	0.404	-0.122	-0.037

Table 9: Correlation analysis of model prediction with SCD task, polysemy (#Synsets), and word frequency using Spearman's rank correlation on SemEval En.

ing/enhancing the effect of frequency/polysemy will likely lead to further improvements in SCD performance.

Much prior work on SCD use the Euclidean distance (or cosine similarity) for measuring semantic change scores of words (Laicher et al., 2021; Giulianelli et al., 2022; Rosin et al., 2022; Rosin and Radinsky, 2022; Cassotti et al., 2023), which (a) weights all dimensions equally, and (b) does not consider cross-dimensional correlations. However, our experimental results show that Euclidean distance is a suboptimal choice for this purpose and learning a Mahalanobis distance metric is more appropriate.

6 Conclusion

We proposed a supervised two-staged SCD method to address two challenges in the SCD tasks: 1) models must obtain *sense-aware* embeddings of target words over time, and 2) due to the lack of labelled training data, SCD is often an unsupervised task using conventional distance metrics. For the first challenge, we propose to learn a *sense-aware* encoder. Next, we address the second challenge by learning a *sense-aware* distance metric to compare *sense-aware* embeddings. In both stages, we used WiC datasets to provide sense-level supervision. Experimental results show that our proposed method, SDML, achieves strong performance in four SCD benchmarks. Moreover, SDML achieves significant performance improvement in WiC benchmarks. Our findings highlight the importance of learning both a sense-aware encoder and a sense-aware distance metric.

Limitations

We show that our method can achieve strong performance in three languages (English, German, and Russian). However, due to the limited training datasets, our method cannot perform well in other languages such as Swedish and Latin. A potential solution to address this limitation and to further improve our method for languages not covered by existing WiC training datasets, is to explore the possibility of using cross-lingual language transfer methods.

Ethical Considerations

In this paper, we focus on detecting semantic changes of words over time. We did not create new datasets and used existing datasets for WiC and SCD for training and evaluation. To the best of our knowledge, no ethical issues have been reported related to those datasets. However, we used publicly available and pretrained MLMs in this paper, and some of those MLM are known to encode unfair social biases such as gender or race (Basta et al., 2019). It is possible that some of those social biases will be present (and possibly have been amplified) during the sense-aware encoder training process. Therefore, we consider it to be an important and necessary task to evaluate the sense-aware encoder that we trained for any social biases before it is used in downstream tasks.

References

Taichi Aida and Danushka Bollegala. 2023a. Swap and predict – predicting the semantic changes in words across corpora by context swapping. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 7753–7772. https://doi.org/10.18653/v1/2023.findings-emnlp.520.

Taichi Aida and Danushka Bollegala. 2023b. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, pages 6868–6882. https://doi.org/10.18653/v1/2023.findings-acl.429.

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*. Association for Computational Linguistics, Shanghai, China, pages 21–31. https://aclanthology.org/2021.paclic-1.3.

Nikolay Arefyev, Maksim Fedoseev, Vitaly Protasov, Daniil Homskiy, Adis Davletov, and Alexander Panchenko. 2021. Deepmistake: Which senses are hard to distinguish for a word-in-context model. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*. volume 20. https://doi.org/10.28995/2075-7182-2021-20-16-30.

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*. Association for Computational Linguistics, Florence, Italy, pages 33–39. https://doi.org/10.18653/v1/W19-3805.

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*. International Committee for Computational Linguistics, Barcelona (online), pages 50–58. https://doi.org/10.18653/v1/2020.semeval-1.4.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 590–602. https://doi.org/10.18653/v1/2023.acl-short.52.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 1577–1585. https://doi.org/10.18653/v1/2023.acl-short.135.

Yair Censor and Stravoz A. Zenios. 1997. *Parellel Optimization*. Oxford University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.

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)*. European Language Resources Association (ELRA), Valletta, Malta.

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*. Association for Computing Machinery, New York, NY, USA, ICML '07, page 209–216. https://doi.org/10.1145/1273496.1273523.

Jeffrey Dean and Sanjay Ghemawat. 2004. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*. San Francisco, CA, pages 137–150.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 2069–2075. https://doi.org/10.18653/v1/N19-1210.

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*. Association for Computational Linguistics, Florence, Italy, pages 457–470. https://doi.org/10.18653/v1/P19-1044.

Christiane Fellbaum and George Miller. 1998. *WordNet: An electronic lexical database*. MIT press. https://ieeexplore.ieee.org/book/6267389.

Mario Giulianelli, Andrey Kutuzov, and Lidia Pivovarova. 2022. Do not fire the linguist: Grammatical profiles help language models detect semantic change. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Dublin, Ireland, pages 54–67. https://doi.org/10.18653/v1/2022.lchange-1.6.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*. IEEE, volume 2, pages 1735–1742.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1489–1501. https://doi.org/10.18653/v1/P16-1141.

Zellig Harris. 1954. Distributional structure. *Word* 10(23):146–162.

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*. Association for Computational Linguistics, Baltimore, MD, USA, pages 61–65. https://doi.org/10.3115/v1/W14-2517.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW 2015*. pages 625–635.

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*. International Committee for Computational Linguistics, Barcelona (online), pages 126–134. https://doi.org/10.18653/v1/2020.semeval-1.14.

Andrey Kutuzov, Lilja Ovrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 1384–1397. https://aclanthology.org/C18-1117.

Andrey Kutuzov and Lidia Pivovarova. 2021. Three-part diachronic semantic change dataset for Russian. In Nina Tahmasebi, Adam Jatowt, Yang Xu, Simon Hengchen, Syrielle Montariol, and Haim Dubossarsky, editors, *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*. Association for Computational Linguistics, Online, pages 7–13. https://doi.org/10.18653/v1/2021.lchange-1.2.

Andrey Kutuzov, Lidia Pivovarova, and Mario Giulianelli. 2021. Grammatical profiling for semantic change detection. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, pages 423–434. https://doi.org/10.18653/v1/2021.conll-1.33.

Severin Laicher, Sinan Kurtyigit, 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*. Association

for Computational Linguistics, Online, pages 192–202. https://doi.org/10.18653/v1/2021.eacl-srw.25.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=73OmmrCfSyy.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7151–7162. https://doi.org/10.18653/v1/2021.emnlp-main.571.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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*. Association for Computational Linguistics, Dublin, Ireland, pages 251–260. https://doi.org/10.18653/v1/2022.acl-demo.25.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, Online, pages 24–36. https://doi.org/10.18653/v1/2021.semeval-1.3.

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*. European Language Resources Association, Marseille, France, pages 4811–4819. https://aclanthology.org/2020.lrec-1.592.

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. https://doi.org/10.1126/science.1199644.

Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. 2023. Variance matters: Detecting semantic differences without corpus/word alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 15609–15622. https://doi.org/10.18653/v1/2023.emnlp-main.965.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (chat)GPT v BERT dawn of justice for semantic change detection. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, St. Julian's, Malta, pages 420–436. https://aclanthology.org/2024.findings-eacl.29.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 1267–1273. https://doi.org/10.18653/v1/N19-1128.

Maxim Rachinskiy and Nikolay Arefyev. 2021. Zeroshot crosslingual transfer of a gloss language model for semantic change detection. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*. volume 20. https://doi.org/10.28995/2075-7182-2021-20-578-586.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 7193–7206. https://doi.org/10.18653/v1/2020.emnlp-main.584.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 3980–3990.

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*. Association for Computing Machinery, New York, NY, USA, WSDM '22, pages 833–841. https://doi.org/10.1145/3488560.3498529.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, pages 1498–1508. https://doi.org/10.18653/v1/2022.findings-naacl.112.

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*. International Committee for Computational Linguistics, Barcelona (online), pages 1–23. https://doi.org/10.18653/v1/2020.semeval-1.1.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 819–862.

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

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, pages 3575–3590. https://doi.org/10.18653/v1/2023.findings-emnlp.231.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Prior and current work on semantic change*, Cambridge University Press, page 51–104. Cambridge Studies in Linguistics. https://doi.org/10.1017/CBO9780511486500.004.

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. https://doi.org/10.1145/3159652.3159703.

Zhao yu Su, Zecheng Tang, Xinyan Guan, Juntao Li, Lijun Wu, and M. Zhang. 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*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pages 6380–6393. https://aclanthology.org/2022.emnlp-main.428.

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*. Association for Computational Linguistics, Shanghai, China, pages 493–502. https://aclanthology.org/2021.paclic-1.52.

Supplementary Materials

A Details of ITML

In this section, we describe the method we use to learn a sense-aware distance metric, $h(\boldsymbol{w}_1, \boldsymbol{w}_2; \mathbf{A})$, that measures the semantic distance between two sense-aware embeddings $\boldsymbol{w}_1 = \boldsymbol{f}(w, s_1; \boldsymbol{\theta})$ and $\boldsymbol{w}_2 = \boldsymbol{f}(w, s_2; \boldsymbol{\theta})$, for w in two sentences s_1 and s_2, respectively. For the ease of reference, we re-write (2) below as (6), defining this distance metric.

$$h(\boldsymbol{w}_1, \boldsymbol{w}_2; \mathbf{A}) = (\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \mathbf{A} (\boldsymbol{w}_1 - \boldsymbol{w}_2) \quad (6)$$

Here, $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a Mahalanobis matrix (assumed to be positive definite) that defines a (squared) distance metric.

To learn the Mahalanobis distance matrix $\mathbf{A}$ in (6), we use training instances (w, s_1, s_2, y) from a WiC dataset for a target word w and two sentences s_1 and s_2, where $y = 1$ indicates that w has the same meaning in both s_1 and s_2, whereas $y = 0$ indicates that w takes different meanings.

We consider two types of constraints that must be satisfied by h as follows. For instances where $y = 1$, the distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is required to be greater than a distance lower bound ℓ such that, $h(\boldsymbol{w}_1, \boldsymbol{w}_2) \geq \ell$. Likewise, for instances where $y = 0$, the distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is required to be lower than a distance upper bound u such that, $h(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq u$. Recall that there exists a bijection (up to a scaling function) between the set of Mahalanobis distances and the set of the equal mean (denoted by $\boldsymbol{\mu}$) multivariate Gaussian distributions given by $p(\boldsymbol{w}; \mathbf{A}) = \frac{1}{Z} \exp\left(-\frac{1}{2} h(\boldsymbol{w}, \boldsymbol{\mu}; \mathbf{A})\right)$, where Z is a normalising constant and $\mathbf{A}^{-1}$ is the covariance of the distribution. We can use this bijection to measure the distance between two Mahalanobis distance functions parametrised by $\mathbf{A}_0$ and $\mathbf{A}$ using the relative entropy (KL-divergence) between the correspond-

ing multivariate Gaussians, given by (7).

$$\mathrm{KL}(p(\boldsymbol{w};\mathbf{A}_0)||p(\boldsymbol{w};\mathbf{A})) = \int p(\boldsymbol{w};\mathbf{A}_0)\log\frac{p(\boldsymbol{w};\mathbf{A}_0)}{p(\boldsymbol{w};\mathbf{A})} \tag{7}$$

Relative entropy is a non-negative and convex function in $\mathbf{A}$. The overall optimisation problem is given by (8).

$$\begin{aligned} \min_{\mathbf{A}} \quad & \mathrm{KL}(p(\boldsymbol{w};\mathbf{A}_0)||p(\boldsymbol{w};\mathbf{A})) \\ \text{subject to} \quad & h(\boldsymbol{w}_1,\boldsymbol{w}_2) \leq u \quad y=1, \\ & h(\boldsymbol{w}_1,\boldsymbol{w}_2) \geq \ell \quad y=0. \end{aligned} \tag{8}$$

The optimisation problem given in (8) can be expressed as a particular type of Bregman divergence, which can be efficiently solved using the Bregman's method (Censor and Zenios, 1997). We use the ITML algorithm proposed by Davis et al. (2007) to learn $\mathbf{A}$, which is described in Algorithm 1. We arrange the sense-aware representations $\boldsymbol{f}(w,s_1)$ and $\boldsymbol{f}(w,s_2)$ for each instance (w,s_1,s_2,y) as columns to create the input matrix $\mathbf{X} \in R^{d\times 2n}$, where d is the dimensionality of the sense-aware embeddings produced by the encoder trained in § 3.1, and n is the total number of training instances in the WiC dataset. The initial value of the distance matrix, $\mathbf{A}_0$ is set to the identity matrix, which corresponds to computing the Euclidean distance. The upper bound, u, is set to the distance that covers the top 5% of the distances between positive instances (i.e. $y=1$), while the lower bound, l, is set to the distance that covers the bottom 5% of the distances between the negative instances ($y=0$).

Algorithm 1 Information Theoretic Metric Learning (ITML)

Input: input matrix $\mathbf{X}$, labels $\boldsymbol{y}$, distance thresholds $[l,u]$, input Mahalanobis matrix $\mathbf{A}_0$, slack parameter γ
Output: Mahalanobis matrix $\mathbf{A}$

```
# Initialise A, λ, and ξ
A ← A_0
for i = 1 to n do
    λ_i ← 0
    ξ_i ← u if y_i = 1 else l
end for
# Optimise A
repeat
    for i = 1 to n do
        obtain i-th instance (w_1, w_2, y_i) from X and y
        d ← h(w_1, w_2; A) in (6)
        δ ← 1 if y_i = 1 else −1
        α ← min(λ_i, δ(1/d − γ/ξ_i)/2)
        ξ_i ← γξ_i/(γ + δαξ_i)
        λ_i ← λ_i − α
        β ← δα/(1 − δαd)
        A ← A + βα(w_1 − w_2)(w_1 − w_2)^⊤ A
    end for
until convergence
return A
```

B Data Statistics

We provide the full data statistics for the SCD benchmarks in Table 4.

Dataset	Language	Time Period	#Targets	#Sentences	#Tokens	#Types
SemEval-2020 Task 1	English	1810–1860	37	254k	6.5M	87k
		1960–2010		354k	6.7M	150k
	German	1800–1899	48	2.6M	70.2M	1.0M
		1946–1990		3.5M	72.3M	2.3M
	Swedish	1790–1830	31	3.4M	71.0M	1.9M
		1895–1903		5.2M	110.0M	3.4M
	Latin	B.C. 200–0	40	96k	1.7M	65k
		0–2000		463k	9.4M	253k
RuShiftEval	Russian	1700–1916	99	3.3k	97k	39k
		1918–1990		3.3k	78k	34k
		1992–2016		3.3k	78k	35k

Table 10: Full statistics of the SCD datasets.