Fake it Till You Make it: Self-Supervised Semantic Shifts for Monolingual Word Embedding Tasks

Maurício Gruppi,[1] Pin-Yu Chen[2] Sibel Adalı,[1]
[1] Rensselaer Polytechnic Institute, NY, USA
[2] IBM Research, NY, USA
gouvem@rpi.edu, pin-yu.chen@ibm.com, adalis@rpi.edu

Abstract

The use of language is subject to variation over time as well as across social groups and knowledge domains, leading to differences even in the monolingual scenario. Such variation in word usage is often called lexical semantic change (LSC). The goal of LSC is to characterize and quantify language variations with respect to word meaning, to measure how distinct two language sources are (that is, people or language models). Because there is hardly any data available for such a task, most solutions involve unsupervised methods to align two embeddings and predict semantic change with respect to a distance measure. To that end, we propose a self-supervised approach to model lexical semantic change by generating training samples by introducing perturbations of word vectors in the input corpora. We show that our method can be used for the detection of semantic change with any alignment method. Furthermore, it can be used to choose the landmark words to use in alignment and can lead to substantial improvements over the existing techniques for alignment. We illustrate the utility of our techniques using experimental results on three different datasets, involving words with the same or different meanings. Our methods not only provide significant improvements but also can lead to novel findings for the LSC problem.

1 Introduction

Language use is deeply rooted in the social, cultural and historical context that shapes it. It has been shown that meaning of words change over time, being pushed by cultural and societal transformations, a phenomenon named *semantic change* (Schmidt 1963). For example, the English word *awful* was used in the sense of the words *impressive* or *majestic* before the year 1800, while, in modern English, it describes *something objectionable*. Language is also subject to variation across different communities for many different reasons (Schlechtweg et al. 2019). For example, the word *model* can be used to refer to the design or version of a product (as in model of a car), or it could be used to refer to a mathematical model in a scientific paper.

In this paper, we develop a novel self-supervised semantic shift (S4) method to detect words with different meaning in two corpora from the same language (monolingual).

As methods based on the distributional property of words have been shown to be very effective in encoding semantic relationship between words (Hamilton, Leskovec, and Jurafsky 2016b; Sagi, Kaufmann, and Clark 2009; Bamler and Mandt 2017; Kulkarni et al. 2015) or even biases and stereotypes (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017), the task of identifying semantic change between words using word embeddings, such as Word2Vec or FastText (Mikolov et al. 2013; Bojanowski et al. 2017) has gained a great deal of popularity. This task is often a difficult one as it involves unsupervised methods (e.g. learning embeddings, alignment and/or mapping of words). For example, in the recent SemEval-2020 competition (Schlechtweg et al. 2020), the highest scores were at about 70% accuracy on a binary classification task to predict occurrence of semantic change across time periods in several languages. The main challenge stems from the unsupervised nature of the problem, as training data is rare or non-existing, and is highly dependent on the input corpora. This impacts multiple aspects of the task. In particular, to compare the embedding matrices $A, B \in \mathbb{R}^{n\times d}$ of two separate corpora $\mathcal{A}$ and $\mathcal{B}$, where n is the size of the common vocabulary and d is the embedding dimension, one must first align them to make them comparable, usually via an Orthogonal Procrustes (OP) method. The goal of OP is to learn an orthogonal transform matrix Q (i.e., $Q^TQ = I_d$, where I_d is the d-dimensional identity matrix), that most closely maps A to B, namely, $Q^* = \arg\min_{Q:Q^TQ=I_d} \|AQ - B\|$. It has been shown that this problem accepts a closed-form solution via Singular Value Decomposition (SVD) (Schönemann 1966). The fact that Q is an orthogonal matrix makes it so that AQ is only subject to unitary transformations such as reflection and rotation, preserving the inner product between its word vectors. Words whose vectors are used in the OP are called *landmarks* (or anchors), these are the words over which we will enforce proximity in the alignment by minimizing the distance of its vectors.

Any landmark choice incorporates an initial assumption to the solution. An ideal solution to this problem would involve a set of landmark words that are semantically stable (i.e. words that have the same sense) across the input corpora. In the context of diachronic embeddings, where the embedding is learned from adjacent time slices of text, the assumption is that words will change only slightly from time

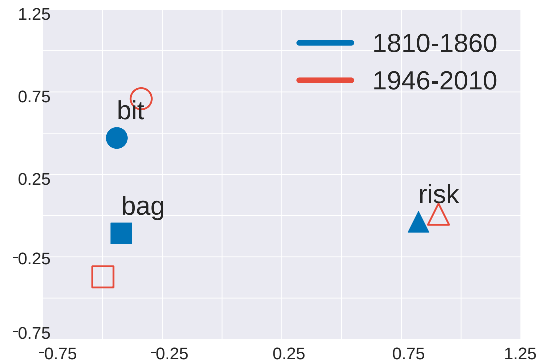
t to $t+1$, therefore, without loss of generality, all words can be used as landmarks. We refer to the use of all common words as landmarks as the global alignment. Global alignment may introduce undesirable similarity between word vectors of a word w that is used in different senses across corpora. As a consequence of the orthogonal transformation, words that are supposed to be closely aligned will be distant. Wang et al. (Wang et al. 2019) refer to this problem, in the context of word translation, as *oversharing*, that is, aligning more words than one should.

This problem is illustrated in Figure 1. We aligned the embeddings of English historical texts from the 19^{th} and 20^{th}-21^{st} centuries. The data and ground-truth in this example was extracted from the SemEval-2020 task on unsupervised lexical semantic change detection (Schlechtweg et al. 2020). The word 'bit' is known to have suffered semantic change over time, it has received the new sense of binary digit. Words 'bag' and 'risk' are labelled as semantically stable over the given period. We show the results of Global Alignment for this data in Figure 1(a) and alignment with landmarks chosen with our proposed method S4-A in Figure 1(b). In Figure 1(a), by doing Global Alignment, we are forcing 20^{th} century *bit* (red circle) to be close to 19^{th} century *bit* (blue circle) and also 20^{th} century *bag* (red square) to be further away from the 19^{th} century *bag* (blue square). Notice in (b) that by choosing landmarks using S4-A, words *bag* and *risk* from two different time periods remain close, while bit from the different time periods is far, hence amplifying the difference between semantically stable and unstable words. More specifically, the cosine distance across the time periods for *bit* is 0.40, and 0.41 for *bag* in (a). Those respective quantities are 0.61 and 0.01 in (b).

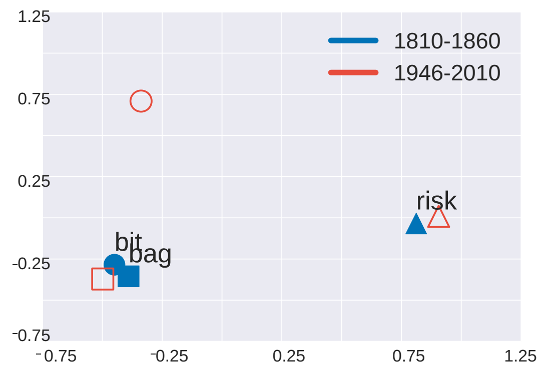
Self-supervised techniques have been shown to be effective in similarly challenging vision-based tasks with limited (or even w/o) supervision (Dosovitskiy et al. 2015; Doersch, Gupta, and Efros 2015; Zhang, Isola, and Efros 2016; Gidaris, Singh, and Komodakis 2018), by using techniques such as image rotation and context changes to create pseudo labels for learning. In this paper, motivated by these advances, we introduce a new method we call *self-supervised semantic shift* (S4 for short) to guide the self-supervision by adding perturbations to set of target word vectors by moving them towards other words in the vector space. This process mimics the occurrence of semantic change in the corpora, making specific words semantically unstable.

This paper makes the following contributions:

1. We introduce the novel S4 method and demonstrate its advantage on two tasks in monolingual word embedding: Unsupervised Binary Classification, and Word Embedding Alignment. In classification, S4 simulates semantic changes and learns to predict semantically stable and unstable words via self-supervision. In alignment, S4 learns to determine the landmark set by predicting, selecting, and refining a set of stable words with self-supervision.
2. We evaluate S4-D's classification performance on a British v.s. American English semantic change detection task. Regardless of the underlying alignment methods, S4 consistently shows significant gain over the baseline methods, attaining up to 2.65× higher F1 score.
3. For the SemEval-2020 task (Schlechtweg et al. 2020), we show landmark words learned by S4-Aattains improved performance on lexical semantic change detection (with a simple linear classifier) in four different languages, demonstrating the importance of landmark selection with S4-A for downstream tasks.
4. We also use S4 for discovery of semantic changes in articles within two arXiv subjects: Artificial Intelligence and Classical Physics. We find that S4-based alignment can identify unique semantically changed words in top-ranked word lists that were overlooked by existing alignment methods, providing diversity and novel discovery of lexical semantic changes.



(a) Global Alignment.



(b) Landmarks selected with S4-A.

Figure 1: PCA projection of aligned word vectors using: (a) the global alignment strategy, (b) landmarks selected with S4-A. Blue (and red resp.) marks show the location of a word in 1810-1860 (and 1946-2010 resp.). Only the word 'bit' is subject to semantic change in this example. Shapes correspond to words *bit* (circle), *bag* (square), *risk* (triangle). Cosine distances of *bit*, *bag*, and *risk* across time periods in are 0.40, 0.41 and 0.01, and 0.61, 0.01, and 0.01 in (b), respectively.

Our methods have many applications in which semantic change has been shown to be important such as diachronic linguistic analysis (Hamilton, Leskovec, and Jurafsky 2016a; Dubossarsky, Weinshall, and Grossman 2017), and predicting armed conflict participants from semantic relations (Kutuzov, Velldal, and Øvrelid 2017). More recently, Bojanowski et al. (2019) have presented a strategy using alignment of monolingual embeddings for updating language models to incorporate changes caused by language evolution or usage in specific domains.

Code for the experiments in this paper are provided in our GitHub repository [1].

2 Related Work

Some of the earliest works on diachronic semantic change discovery analyzes word usage over time with respect to frequency and sense, mostly based on the difference of the distributional property of words between time periods or domains (Sagi, Kaufmann, and Clark 2009; Cook and Stevenson 2010). Distributed word vector representations, such as the ones obtained by skip-gram with negative sampling (Mikolov et al. 2013) allow for learning distributional information into dense continuous vectors. Hamilton et al. (2016b) conducted a diachronic analysis of semantic change on historical text using word embeddings aligned with Orthogonal Procrustes and by measuring the displacement of vectors across time periods using cosine distance. To circumvent the need for alignment, Hamilton et al. (2016a) have proposed a measure of semantic change that compares second-order vectors of distance between a word and its neighbors. Some authors have presented dynamic word embedding techniques to avoid alignment by jointly learning the distributional representations across all time periods (Bamler and Mandt 2017; Rudolph and Blei 2018; Yao et al. 2018), in which words are connected across time via the assumption that the change across periods is smooth. Yin et al. (Yin, Sachidananda, and Prabhakar 2018) introduced the Global Anchor method for corpus-level adaptation, which avoids alignment altogether by using second-order distances. This method is proven to be equivalent to the global alignment method and as a result makes use of the smoothness of change assumption which may lead to oversharing, especially in cross-domain scenarios.

Selection of landmarks as words with likely same meaning in two different languages is used in translation tasks. Artetxe et al. (2017) employ a self-supervision approach to refine a small seed dictionary with arabic numerals. Conneau et al. (2017) use self-supervising adversarial learning to align bilingual embeddings, refining it with Orthogonal Procrustes on the best matching words. Joulin et al. (2018) propose an alternative loss function for the alignment in order to address the conflict between Euclidean alignment and cosine distance mapping and to refine the alignment by matching words that have similar frequency ranking in each language. Lubin et al. (2019) employ a landmark selection by detecting noisy pairs through an iterative EM algorithm.

Most of the previously developed methods do not present a systematic way of detecting semantic change such as in a classification problem. Our proposed method is designed explicitly for lexical semantic change tasks by matching the same word in two corpora, approaching semantic change as a binary classification problem. Our contributions are: the introduction of a self-supervised method for binary semantic change detection, a method for selecting landmark (anchor) words for alignment of semantically changed word vectors, a quantitative test set on British vs. American English for detecting words with similar and distinct senses. We compare our method both to baseline global alignment and noisy pairs methods and show that it provides gains in performance in a number of scenarios.

3 Self-Supervised Semantic Shift (S4)

The problem of detecting semantic change using monolingual alignment of word embeddings is defined as follows. Given two input corpora $\mathcal{A}$ and $\mathcal{B}$, with vocabularies $V_{\mathcal{A}}$ and $V_{\mathcal{B}}$, let A and B be the word embedding matrices for the common vocabulary $V = V_{\mathcal{A}} \cap V_{\mathcal{B}}$, thus both A and B have dimensions $N \times d$, where N is the size of the common vocabulary, and d is the embedding dimension. A word in the common vocabulary is said to be unstable if it is used in a completely different sense between the corpora, or that has multiple senses but some of which only appear in one corpus, other words are considered stable. One common method for measuring semantic change is to use the cosine distance between two embeddings after aligning A and B on a subset of V. This problem involves two subtasks: detecting words with semantic change and choosing landmark words to align on. In this paper, we introduce a self-supervised method that can be used in both tasks.

Given embeddings A and B, the main goal of the self-supervision is to create a modified embedding B' such that B' contains a set of words that are known to be semantically shifted with respect to their meaning in B through explicit perturbations. We can generate (pseudo) training samples for a self-supervised procedure by using these modified embeddings and the fact that they are considered semantically shifted. Suppose t be the target word whose sense we want to add to another word w. We can accomplish this by replacing t with w an arbitrary number of times r in $\mathcal{B}$ and re-training the word embeddings, where the parameter $r \in (0, 1)$ defines the proportion of replacements with respect to the number of occurrences of t. To reduce complexity, instead of actually retraining the word embeddings, we move the vector v_w towards v_t by the rule $v_w \leftarrow v_w + rv_t$. This update rule is derived from the skip-gram with negative sampling model (Mikolov et al. 2013), specifically, whenever word t occurs within a neighborhood window of w, vector v_w is updated with $v_w + (1 - z)\eta v_t$, where $z = \sigma(v_w^\intercal v_t)$. In our perturbation, we replace $\sigma(v_w^\intercal v_t)$ with parameter r, which is used to control the proportion in which word t is found in the neighborhood of w. The resulting word vector for w in $\mathcal{B}'$ is now shifted towards that of t and w is now forced to become unstable regardless of its original state. By applying this process to multiple words, we are able to generate positive samples (semantically changed words) to the self-supervision. In contrast, negative samples are drawn from the set of landmarks.

[1] S4 code repository: https://github.com/IBM/S4_semantic_shift

Algorithm 1: Pseudo-code for self-supervised semantic shift detection (S4-D). Input parameters are word embeddings A and B, landmark words set L, non-landmark words set M, n as the number of negative and positive samples in each iteration, and r is the degree of perturbation. The output of this method is the classifier weights W.

Data: $A, B, L, M, n, r, max\_iters$
Result: Classifier weights W

$W \leftarrow init\_weights$;
$i \leftarrow 0$;
while $i < max\_iters$ **do**
  $i \leftarrow i + 1$;
  // Sample negatives from L and positives from M
  $S_n = uniform\_sample(L, n)$;
  $S_p = uniform\_sample(M, n)$;
  $B' \leftarrow copy(B)$;
  for $w \in S_p$ **do**
    $t \leftarrow uniform\_sample(M)$;
    // Simulate change by moving w towards t
    $B'(w) \leftarrow B(w) + rB(t)$;
  end
  $X \leftarrow [A(w), B'(w)]\ \forall w \in S_n \cup S_p$;
  $Y \leftarrow [0 \text{ if } w \in S_n \text{ else } 1]\ \forall w \in S_n \cup S_p$;
  $W \leftarrow train(W, X, Y)$;
end
return W

3.1 S4-D: S4 for Semantic Change Detection

Self-supervision for semantic change detection can be used in conjunction with any method that uses a subset of words as landmarks. Given an initial alignment of A to B on a set of landmarks $\mathcal{L}$ (potentially using Orthogonal Procrustes), a batch of positive samples is generated from the perturbations, and negative samples are uniformly drawn from $\mathcal{L}$. These samples are then used to train a binary classifier to predict stable vs. unstable words. We use a single-layer neural network classifier with 100 hidden units with ReLU activation and sigmoid output. The input to the model is the concatenation of row vectors $A(w)$ and $B(w)$ for word w. The model is trained over a predefined number of iterations K to predict $\hat{y} = 0$ if w is stable, otherwise $\hat{y} = 1$. A new batch of positive and negative samples is generated in every iteration. The goal is to minimize the average loss $\frac{1}{K}\sum_{i=1}^{K} L(\hat{y}_i, y_i)$, with L as the binary cross-entropy function.

Note that, at this point, the self-supervision is done over a fixed alignment of A and B and it is trained to predict semantic change on that setup. The hyper-parameters of S4-D are: number of iterations K, number of negative and positive samples to generate in each iteration n and m, degree of semantic change in the perturbations r. The pseudo-code for S4-D is presented in Algorithm 1.

3.2 S4-A: S4 for Alignment of Word Vectors

In this section, we present an extension of the self-supervised training to refine the landmarks based on the classifier predictions from Section 3.1, resulting in the Self-Supervised Semantic Shift Alignment (S4-A). The general idea is to use stable words for alignment by adding an extra step to each iteration in Algorithm 1. At the end of each training iteration, we update the classifiers weights W and use the updated model to predict stable/unstable words across A and B, hence updating the set of landmarks $\mathcal{L}$ with words predicted as stable. Finally, we align A to B using the new set of landmarks with orthogonal procrustes and repeat over K iterations. This method outputs both model weight and the set of landmark words $\mathcal{L}$ and the set of non-landmark words $\mathcal{N}$. Using the final set $\mathcal{L}$ of landmarks, we can align A to B using orthogonal procrustes on the words in $\mathcal{L}$. Appendix A.1 contains the pseudo-code for this algorithm.

4 Experiments

4.1 Semantic Change Detection

Objective We evaluate the ability of S4-D to correctly detect the occurrence of lexical semantic change across British and American English. We designed binary classification task to evaluate the model's performance in predicting semantically stable vs. unstable words. The full list of words is given in Appendix B.

Data set The corpus used for British English is the British National Corpus (BNC) XML Edition (of Oxford 2007) which contains a mix of news, fiction, and academic texts. The corpus for American English is the Corpus of Contemporary American English (COCA) (Davies 2009) which contains text from newspapers, magazines, fiction, and academic texts. Both corpora are pre-processed by removing stop words, and converting all characters to lower case, resulting in 50M tokens for BNC and 188M tokens for COCA.

Baselines The baseline methods used are the commonly used cosine distance-based method (Hamilton, Leskovec, and Jurafsky 2016b; Kutuzov et al. 2018; Schlechtweg et al. 2019), and the Noise-Aware method (Yehezkel Lubin, Goldberger, and Goldberg 2019), which detects noisy word pairs. For the cosine distance (COS), we use three different thresholds for this measure, specifically we have three cosine-based classifiers at thresholds 0.3, 0.5, and 0.7 , above which words are classified as semantically shifted. For Noise-Aware, we treat noisy word pairs as semantically shifted (unstable), and clean word pairs as stable, we will refer to that method as *Noisy-Pairs*. We compare the baseline methods to our proposed Self-Supervised Semantic Shift Detection (S4-D) with hyper-parameters $n = 1000$ positives samples, $m = 1000$ negative samples, rate $r = 0.25$, trained over 100 iterations (at each iteration, a batch of 1000 positive and 1000 negative samples are generated and used to update the model's weights). We train Word2Vec on the input corpora after pre-processing with parameters dimension 300, window size 10, minimum count 100 for British, and dimension 300, window size 10, minimum count 200 for American.

The minimum count for the American corpus is set higher due to its corpus being considerably larger. The common vocabulary contains 26064 words.

Evaluation and Analysis Once the embeddings X_A, X_B are learned from the COCA and BNC corpora, respectively, we get the embedding matrices for the common vocabulary $A, B \in \mathbb{R}^{n \times d}$ by selecting the rows of X_A and X_B that correspond to words in the common vocabulary. We learn a transform matrix Q by aligning A to B using a given alignment strategy. The self-supervised classifier is trained on A and B using the given alignment. Using the learned matrix Q, we align X_B to X_A by doing $X_B \leftarrow X_B Q$. Finally, we concatenate the vectors of the target words and feed it into the classifier to obtain the binary predictions (i.e. semantically stable or unstable).

The classification scores for this task (Table 1) show that S4-D displays the best accuracy when aligning on the top 10% most frequent words, it also shows high recall and F1 scores when aligning on the 10% least frequent and 5% most frequent words. The scores for S4-D are the average of 10 evaluation rounds, standard deviation is also reported in Table 1. Each evaluation round consists of one execution of the algorithm on the input data. This contrasts with the drop in performance shown by global alignment, this is likely due to the oversharing of words, which makes the separation of stable and unstable words more difficult. The alignment method is irrelevant to Noisy-Pairs since it inherently aligns the input vectors when searching for noise. Noisy-Pairs predicts a total of 24659 pairs as clean, or semantically stable. For that reason, only one pair of words from the target set is predicted as semantically shifted (positive class), which explains the precision score of 1.0 and a recall of 0.03.

Examples of stable words correctly predicted by S4-D are the British-American pairs *labour/labor*, *defence/defense*, *petrol/gas*, *football/soccer*, and *queue/line*. This shows we are able to not only to detect identical words but also morphological differences and synonyms. Note that some of these words were not included in the alignment due to not being in the common vocabulary. Yet, we are still able to capture their semantic similarity after the orthogonal transformation. These results show our model's ability to generalize to words not seen in the self-supervision. Additionally, we were able to correctly predict unstable words such as *chips* (french fries in the US), *biscuit* (scone in the UK, cookie in the US), and *semi* (house in the UK, truck in the US). Noisy-Pairs is able to correctly predict the semantically unstable words *subway* and *yankee*, and it also predicts all stable words correctly but shows a low recall score.

4.2 Evaluating Alignment Strategies

Objective We evaluate the Self-Supervised Semantic Shift Alignment (S4-A) described in Section 3.2 by the impact of assessing multiple alignment strategies on the performance of the binary classification problem from the recent SemEval-2020 task on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al. 2020). The task consists of predicting the occurrence of lexical semantic change on a set of target words in four languages: English, German, Latin, and Swedish. For each language there are two input corpora C_1 and C_2 containing text from time periods t_1 and t_2, with $t_1 < t_2$.

Data sets The data sets used in this experiment are provided in the aforementioned SemEval task. The corpus used for English is the Clean Corpus of Historical American English (CCOHA) (Alatrash et al. 2020), a pre-processed and lemmatized version of the Corpus of Historical American English (COHA) (Davies 2009). For this task, the corpus was split into time periods t_1 with texts from years 1810 through 1860, and t_2 with text from years 1960 through 2010. The German data set consists of the DTA corpus (von der Berlin-Brandenburgischen Akademie der Wissenschaften 2017) for the first time period, and a combination of the BZ and ND corpora (zu Berlin 2018a,b) for the second time period. Specifically, text in t_1 pertains to years 1800 through 1899, and text in t_2 pertains to years 1946 through 1990. For Latin we use the LatinISE corpus (McGillivray and Kilgarriff 2013) with time periods t_1 from 200 B.C. through 0 A.D., and t_2 from years 0 through 2000 (A.D.). For Swedish we use the KubHist corpus (Borin, Forsberg, and Roxendal 2012; Adesam, Dannélls, and Tahmasebi 2019). Time period t_1 is from 1790 through 1830, time period t_2 is from 1895 through 1903. Along with each data set, a list of target words is provided for evaluation. Further details about the data sets can be found in Appendix C.

Baselines We include common alignment methods for comparison to S4-A. Particularly, we adopt the global alignment strategy as used in diachronic embeddings (Hamilton, Leskovec, and Jurafsky 2016b), as well as frequency based selection of landmarks, aligning at the 5% and 10% most and least frequent words (Bojanowski et al. 2017), and the Noise-Aware method for selecting landmarks.

Evaluation and Analysis We begin by training Word2Vec on C_1 and C_2, generating embeddings X_1 and X_2. We set the embedding dimension to 300 and use a window of size 10 for all languages. The minimum word count is 20, 30, 10, and 50 for English, German, Latin, and Swedish, respectively, these are chosen based on the amount of data provided for each language.

Let $A \subset X_1$ and $B \subset X_2$ be the embedding matrices for the common vocabulary terms in C_1 and C_2. Our experiment consists of aligning A to B using different alignment strategies and evaluating the alignments with respect to its performance in the binary classification task. Particularly, we evaluate our self-supervised alignment and compare it to doing global alignment, aligning on most and least frequent words, and selecting clean words as landmarks (Yehezkel Lubin, Goldberger, and Goldberg 2019).

Since there is no labeled training data for this problem, we build a model to predict lexical semantic change based on the cosine distance between the word vectors in A and B, after alignment. We compute the cosine distance between all pairs of vectors in A and B. Then, for each word w in the common vocabulary, we compute the cumulative probability $P(x < X)$ where x is the cosine distance between the vectors of w. Finally, we decide on the class of w based on a threshold $t \in (0, 1)$. To determine the value of t, we perform model selection through cross-validation on the self-supervised data, selecting t that achieves the best accuracy

Method	Alignment	Accuracy	Precision	Recall	F1
COS	Global	$0.38/0.30/0.25$	$0.48/0.36/0.00$	$0.26/0.10/0.00$	$0.32/0.15/0.00$
S4-D		0.45 ± 0.03	0.81 ± 0.06	0.28 ± 0.05	0.43 ± 0.06
COS	Top-5%	$0.39/0.30/0.25$	$0.47/0.36/0.00$	$0.29/0.10/0.00$	$0.33/0.15/0.00$
S4-D		0.67 ± 0.02	0.81 ± 0.02	0.79 ± 0.02	$\mathbf{0.82} \pm 0.01$
COS	Top-10%	$0.37/0.30/0.25$	$0.47/0.36/0.00$	$0.26/0.10/0.00$	$0.31/0.15/0.00$
S4-D		$\mathbf{0.70} \pm 0.03$	$\mathbf{0.83} \pm 0.02$	0.74 ± 0.03	0.78 ± 0.02
COS	Bot-5%	$0.40/0.29/0.24$	$0.46/0.32/0.00$	$0.31/0.10/0.00$	$0.35/0.16/0.00$
S4-D		0.45 ± 0.02	0.62 ± 0.01	0.21 ± 0.02	0.31 ± 0.01
COS	Bot-10%	$0.37/0.30/0.25$	$0.53/0.45/0.25$	$0.29/0.12/0.02$	$0.34/0.18/0.03$
S4-D		0.65 ± 0.02	0.71 ± 0.01	0.89 ± 0.02	0.79 ± 0.01
COS	S4-A	$0.44/0.34/0.28$	$0.66/0.62/0.57$	$0.40/0.18/0.06$	$0.44/0.27/0.12$
S4-D	S4-A	$\mathbf{0.70} \pm 0.01$	0.72 ± 0.01	$\mathbf{0.93} \pm 0.01$	0.81 ± 0.01
Noisy-Pairs	-	0.30	$\mathbf{1.00}$	0.03	0.06

Table 1: Classification scores for the British vs. American English task. The baseline cosine (COS) method outputs unstable words whose cosine distance is greater than $0.3/0.5/0.7$. Top-N/bot-N alignments use the most/least frequent words. S4-D (our method) is the self-supervised semantic shift detection. The scores for S4-D are given as the mean and standard deviation over 10 evaluation rounds. The initial alignment is irrelevant to noisy-pairs as it necessarily searches an alignment.

in the leave-one-out tests, t is searched in $(0, 1)$ in increments of 0.1. The prediction is $\hat{y} = 1$ if $P(x < X) > t$, otherwise $\hat{y} = 0$.

Results from this experiment are shown in Table 2. We report the accuracy on the evaluation set for each language and alignment method. Top/bot alignments are done over the top 5%/10% most/least frequent words in C_1. S4-A is able to achieve the best accuracy scores maximum scores for English and German, matching some of the top performing scores in the post-evaluation phase of the SemEval-2020 Task 1 competition [2]

4.3 Discovery of Semantic Change

Objective We conduct an experiment on the arXiv data provided by Yin et al. (2018) to show how we can use S4-A for word embedding alignment for the discovery of semantic change, and how the results differ across alignment methods. We select the subjects of Artificial Intelligence (cs.AI) and Classical Physics (physics.class-ph) and train embeddings A and B, respectively, with Word2Vec (dimension 300, window size 10, minimum count 20). The embedding matrices are aligned using each alignment strategy, and the semantic shift measured by $d_i = \|A_i Q - B_i\|$ for each word w_i in the common vocabulary, where Q is the transform matrix learned in the alignment.

Baselines We compare the most semantically shifted words as discovered by the Global and Noise-Aware alignments (Hamilton, Leskovec, and Jurafsky 2016b; Yehezkel Lubin, Goldberger, and Goldberg 2019). We also compare our results to the top 3 high scoring entries from post-evaluation phase of the SemEval-2020 Task 1 competition, these methods may use distinct sets of features that go beyond just using word embeddings.

[2] https://competitions.codalab.org/competitions/20948#results Accessed: 2020-08-04.

Alignment	English	German	Latin	Swedish
S4-A	**0.70**	**0.81**	0.68	0.77
Noise-Aware	0.65	0.79	0.65	0.74
Top 5% fr.	0.65	0.77	0.68	0.77
Top 10% fr.	0.68	0.79	0.68	0.74
Bot 5% fr.	0.68	0.73	0.62	0.77
Bot 10% fr.	0.68	0.75	**0.70**	**0.81**
Global	0.68	0.79	0.65	0.74
SemEval #1	**0.70**	0.79	0.68	**0.81**
SemEval #2	**0.70**	0.77	0.72	0.74
SemEval #3	0.65	0.77	0.75	0.74

Table 2: Classification accuracy of the unsupervised lexical semantic change detection on the SemEval-2020 Task 1 data set. The results were obtained by aligning the embedding matrices using different alignment strategies, and applying a threshold to the cosine distance of the aligned vectors, selected by cross-validation. Top-fr. and Bot-fr. are alignments using OP on the top and bottom 5% and 10% frequent words. The bottom rows shows the top 3 high scoring submissions to SemEval-2020 Task 1 in the post evaluation phase.

Evaluation and Analysis To quantify the difference between different alignments, we measured the ranking correlation using the Spearman's rho coefficient of the ranked list of words according to each method (ranked in descending order of semantic shift) at varying top-K thresholds with k in $[10, 500]$ in increments of 10. Figure 2 shows the ranked correlation coefficient between each alignment strategy. Higher values of rho indicate that the order of semantic shift is more consistent between the two alignment strategies. These results reveal that Global and Noise-Aware produce very similar rankings, with rho approaching 1 even for small values of k. On the other hand, the ranking correlation between S4-A is substantially lower for small values of k. This suggests that most of the difference in ranking between S4-A and the others is in the most shifted words, with the ranking of the remaining words being very similar to Global and Noise-Aware. In summary, S4-A can be used to find novel shifted words that are overlooked by existing methods such as Global and Noise-Aware. Table 3 shows the list of uniquely discovered words among the top most shifted for Global and S4-A . Noise-Aware because it does not show any novel words when compared to Global, i.e., its predictions are the same between arXiv subjects of Artificial Intelligence and Classical Physics. We find that words uniquely discovered by S4-A can be naturally explained in context of their subjects, for instance, *mass* is likely more often used as probability mass in AI, and as physical mass in classical physics. More comparisons are shown in Appendix D.

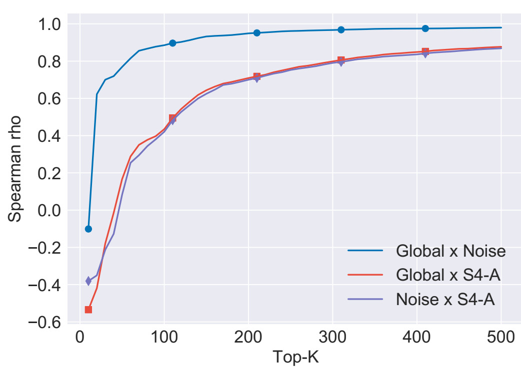


Figure 2: Ranking correlation between Global, Noise-Aware, and S4-A alignments at varying top-k levels. Works are ranked from most to least shifted according to each method and the ranking correlation is measured with Spearman's rho. The semantic shifts are between arXiv subjects cs.AI and physics.class-ph.

Global/Noise-Aware	**S4-A**
agent	component
approximation	element
boundary	mass
conceptual	order
knowledge	solution
plane	space
reference	term
rules	time
system	vector

Table 3: Unique words discovered by each alignment method among the top 50 most shifted, between arXiv cs.AI and physics-class-ph corpora. Global and Noise-Aware show the same predictions in the top 50 words.

5 Conclusions

We introduced S4-D and S4-A as self-supervised approaches to detect word-level semantic shifts on monolingual corpora. Motivated by the unsupervised nature of this problem, we introduce self-supervision based on the perturbation of word vectors and apply it to binary classification and vector alignment. S4-D is presented as an alternative to baseline unsupervised methods for semantic shift detection, particularly in the case of binary classification. We show, through experiments in Section 4.1, that it achieves over $2\times$ higher F1-scores than baselines in the classification settings. Moreover, we show how the alignment of word embeddings affect the outcome of such methods. Particularly, we show that global alignment uses the assumption of smooth transition, which may not hold true in the scenario of cross-domain semantic shift, where many words can be highly shifted. For that reason, we present an extension of our method, named S4-A, that uses its predictions to refine the alignment of the input embeddings. We demonstrate its usefulness quantitatively, through the detection task in Section 4.2, where S4-A allows for the detection of unique words when using a simple cosine distance baseline. Qualitatively, we demonstrate that S4-A is able to discovery novel shifts when compared to other alignment methods.

There are still open questions on how the self-supervised model is affected according to part-of-speech, frequency range, and degree of polysemy of words. In addition, factors such as number of tokens, vocabulary size, and degree of change of the language may impact the quality of the embeddings, therefore, affect the semantic shift detection.

While this remains a difficult task, we believe that this work will help numerous applications of semantic shift detection and alignment that have been recently explored, especially in the monolingual and cross-domain setting.

Ethical Impact

Prior work has shown that embeddings in a single corpus can encode many cultural stereotypes such as gender and racial bias. This is not particularly surprising as language is a tool for creating common meaning. Stereotypes, gender and racial bias are social constructs that have been quite prominent in many cultural artifacts including languages. Code words and dog-whistles are words often used to mean different things only for a specific community. A word that looks quite common and benign can be very offensive when evaluated in context. Some of these differences can be rooted in social justice work of reclaiming controversial words. In short, language codes rich and complex cultural and historical differences that are lost when treated as a monolith. Our emphasis is in creating tools not to smooth over these differences, but help identify them. To accomplish this, we use two corpora in the same language to understand which words in one corpus have a different meaning compared to another corpus. Our work leverages potential bias in languages by using one corpus as a point of reference for the other to highlight the differences.

Acknowledgements

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons). We also thank the anonymous reviewers who contributed with our work with their invaluable feedback.

References

Adesam, Y.; Dannélls, D.; and Tahmasebi, N. 2019. Exploring the Quality of the Digital Historical Newspaper Archive KubHist. In *DHN*, 9–17.

Alatrash, R.; Schlechtweg, D.; Kuhn, J.; and Schulte, S. 2020. Clean Corpus of Historical American English. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*. European Language Resources Association (ELRA).

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 451–462.

Bamler, R.; and Mandt, S. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 380–389. JMLR. org.

Bojanowski, P.; Celebi, O.; Mikolov, T.; Grave, E.; and Joulin, A. 2019. Updating Pre-trained Word Vectors and Text Classifiers using Monolingual Alignment. *arXiv preprint arXiv:1910.06241* .

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135–146.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Borin, L.; Forsberg, M.; and Roxendal, J. 2012. Korp — the corpus infrastructure of Spräkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 474–478. Istanbul, Turkey: European Language Resources Association (ELRA).

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

Cook, P.; and Stevenson, S. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *LREC*.

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14(2): 159–190.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Dosovitskiy, A.; Fischer, P.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2015. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence* 38(9): 1734–1747.

Dubossarsky, H.; Weinshall, D.; and Grossman, E. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 1136–1145.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations* .

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, 2116. NIH Public Access.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016b. 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)*, 1489–1501.

Joulin, A.; Bojanowski, P.; Mikolov, T.; Jégou, H.; and Grave, E. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2979–2984. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1330.

Kulkarni, V.; Al-Rfou, R.; Perozzi, B.; and Skiena, S. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, 625–635.

Kutuzov, A.; Øvrelid, L.; Szymanski, T.; and Velldal, E. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Kutuzov, A.; Velldal, E.; and Øvrelid, L. 2017. Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1824–1829.

McGillivray, B.; and Kilgarriff, A. 2013. Tools for historical corpus research, and a corpus of Latin. *New Methods in Historical Corpus Linguistics* 1(3): 247–257.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

of Oxford, U. 2007. The British National Corpus (BNC XML Edition). URL http://www.natcorp.ox.ac.uk/. Accessed: 2020-07-03.

Rudolph, M.; and Blei, D. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, 1003–1011.

Sagi, E.; Kaufmann, S.; and Clark, B. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111. Association for Computational Linguistics.

Schlechtweg, D.; Hätty, A.; Del Tredici, M.; and Schulte im Walde, S. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 732–746. Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1072.

Schlechtweg, D.; McGillivray, B.; Hengchen, S.; Dubossarsky, H.; and Tahmasebi, N. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *To appear in Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona, Spain: Association for Computational Linguistics.

Schmidt, K. H. 1963. Ullmann Stephen. Semantics. An Introduction to the Science of Meaning (Book Review). *Indogermanische Forschungen* 68: 183.

Schönemann, P. H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31(1): 1–10.

von der Berlin-Brandenburgischen Akademie der Wissenschaften, H. 2017. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. URL http://www.deutschestextarchiv.de/. Accessed: 2020-08-15.

Wang, Z.; Xie, J.; Xu, R.; Yang, Y.; Neubig, G.; and Carbonell, J. 2019. Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework. *arXiv preprint arXiv:1910.04708* .

Yao, Z.; Sun, Y.; Ding, W.; Rao, N.; and Xiong, H. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, 673–681.

Yehezkel Lubin, N.; Goldberger, J.; and Goldberg, Y. 2019. Aligning Vector-spaces with Noisy Supervised Lexicon. 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)*, 460–465. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1045.

Yin, Z.; Sachidananda, V.; and Prabhakar, B. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in neural information processing systems*, 9412–9423.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

zu Berlin, S. 2018a. Berliner Zeitung. Diachronic newspaper corpus published by Staatsbibliothek zu Berlin. URL http://zefys.staatsbibliothek-berlin.de/index.php?id=155. Accessed: 2020-08-15.

zu Berlin, S. 2018b. Neues Deutschland. Diachronic newspaper corpus published by Staatsbibliothek zu Berlin. URL http://zefys.staatsbibliothek-berlin.de/index.php?id=156. Accessed: 2020-08-15.