Mining the UK Web Archive for Semantic Change Detection

Adam Tsakalidis[1], **Marya Bazzi**[1,2,3], **Mihai Cucuringu**[1,3],
Pierpaolo Basile[5], **Barbara McGillivray**[1,4]

[1] The Alan Turing Institute, London, United Kingdom
[2] University of Warwick, Coventry, United Kingdom
[3] University of Oxford, Oxford, United Kingdom
[4] University of Cambridge, Cambridge, United Kingdom
[5] University of Bari, Bari, Italy
{atsakalidis, mbazzi, mcucuringu, bmcgillivray}@turing.ac.uk
pierpaolo.basile@uniba.it

Abstract

Semantic change detection (i.e., identifying words whose meaning has changed over time) started emerging as a growing area of research over the past decade, with important downstream applications in natural language processing, historical linguistics and computational social science. However, several obstacles make progress in the domain slow and difficult. These pertain primarily to the lack of well-established gold standard datasets, resources to study the problem at a fine-grained temporal resolution, and quantitative evaluation approaches. In this work, we aim to mitigate these issues by (a) releasing a new labelled dataset of more than 47K word vectors trained on the UK Web Archive over a short time-frame (2000-2013); (b) proposing a variant of Procrustes alignment to detect words that have undergone semantic shift; and (c) introducing a rank-based approach for evaluation purposes. Through extensive numerical experiments and validation, we illustrate the effectiveness of our approach against competitive baselines. Finally, we also make our resources publicly available to further enable research in the domain.

1 Introduction

Semantic change detection is the task of identifying words whose lexical meaning has changed over time. Detecting this temporal variation enables historical and social scientists to study cultural shifts over time (Michel et al., 2011), but it can also have important implications on the performance of models in various NLP tasks, such as sentiment analysis (Lukeš and Søgaard, 2018).

While early theoretical work on semantic change dates back to the previous century (Bloomfield, 1933), the recent availability of historical datasets has made the computational study of the task feasible (Sandhaus, 2008; Michel et al., 2011; Davies, 2012). Past work has demonstrated that semantic change can manifest over decades (Cook and Stevenson, 2010; Mihalcea and Nastase, 2012), years (Yao et al., 2018; Basile and McGillivray, 2018), or even months and weeks (Kulkarni et al., 2015; Tsakalidis et al., 2018).

However, important gaps make progress in the field slow. In particular, there is a relative lack of labelled datasets to study the task over a short time-frame, since most known instances of semantic change took place over centuries or decades. Furthermore, the evaluation of a semantic change detection model is typically performed by manually inspecting a few examples, which can result in unreliable or even non-measurable performance. Finally, on a methodological front, a common practice to measure the semantic shift of words between consecutive time periods is to calculate their displacement error that results from "aligning" word vector representations across these time periods (Hamilton et al., 2016). However, a subset of these words may have actually undergone semantic change and thus trying to align their representations across time is counter-intuitive for the task of semantic change detection, and – importantly – can result in drop in performance. To this end, our work makes the following contributions:

- We release a new dataset for semantic change detection, comprised of word vector representations trained on yearly time intervals of the UK Web Archive (>20TB), along with a list of words with known semantic change, as provided by the Oxford English Dictionary.
- We propose a variant of Procrustes alignment

Proceedings of Recent Advances in Natural Language Processing, pages 1212–1221,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_139

for semantic shift detection, trained on an extremely small number of "anchor words" whose meaning is "stable" across time.

- We illustrate the effectiveness of our approach through extensive experimentation, by also proposing the employment of rank-based metrics for evaluation purposes.

2 Related Work

Early work on semantic change detection relied primarily on the comparison of word frequency and co-occurrence patterns between words at different time intervals (Sagi et al., 2009; Cook and Stevenson, 2010; Gulordava and Baroni, 2011), most often representing a single word based on its context (Mihalcea and Nastase, 2012; Jatowt and Duh, 2014; Basile and McGillivray, 2018). Recently, word embeddings have become the common practice for constructing word representations in NLP (Mikolov et al., 2013). A typical process followed in the context of semantic change is to learn the representations of a word over different time intervals and then compute its shift, by employing some distance metric over the resulting representations (Kim et al., 2014; Hamilton et al., 2016; Del Tredici et al., 2018).

A key issue that results from this process is that the comparison of the same word across different time periods becomes impossible, due to the stochastic process of generating the word vectors (e.g., word2vec). To accommodate that, Kim et al. (2014) proposed the initialisation of the word embeddings at time $t+1$ based on the resulting word representations at time t. Kulkarni et al. (2015) learned a linear mapping between the word representations of the nearest neighbours of a word at different time periods. Hamilton et al. (2016) employed Orthogonal Procrustes (Schönemann, 1966) to map the resulting word representations of the whole vocabulary at time t to their corresponding ones at time $t+1$. Another strand of work focuses on generating diachronic word embeddings (Kutuzov et al., 2018), aiming to learn word representations across time (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Yao et al., 2018; Rudolph and Blei, 2018). However, these are often hard and slow to train under a massive dataset, such as the UK Web Archive. Similarly, the approach by Kim et al. (2014) does not allow for parallel processing of massive historical collections, since the word vectors at time $t+1$ need to be initialised based on the resulting representations at t. Our work is more closely related to Hamilton et al. (2016). However, aligning the vectors of the whole vocabulary at different times can be noisy and is counter-intuitive for the task of semantic change detection. To mitigate this effect, we propose to learn the alignment based only on a few "stable" (from a semantic point of view) words and apply the same transformation to the full vocabulary, leading to more appropriate alignment and, therefore, to more effective detection of semantically shifted words.

Regardless of the methodological approach, an open issue is the evaluation method of such a model. Owed to the lack of large-scale ground-truth datasets, past work has performed the evaluation either on the basis of detecting only a few word cases of semantic change (Cook and Stevenson, 2010; Gulordava and Baroni, 2011; Del Tredici et al., 2018) or by creating an artificial task, such as word epoch disambiguation (Mihalcea and Nastase, 2012). In this work, we propose instead a rank-based approach that can be employed for the evaluation of a semantic change detection model, even with a few positive examples of words whose lexical semantics have changed.

For more information on semantic change detection, the reader is referred to Tang (2018).

3 Methodology

3.1 Task Definition

Let $[W^{(0)}, ..., W^{(|T|)}]$ be word representations of a common (intersected) vocabulary of $|V|$ terms across $|T|$ consecutive time intervals given by $\{[t, t+1], t \in \{0, \ldots, T-1\}\}$, where each t maps to a given year. Our goal is to *find the words whose meaning has changed the most over each of the consecutive time intervals* $[t, t+1]$ (e.g., between [2000, 2001], [2001, 2002], etc.).

Clear cases of words that have undergone semantic change are difficult to find in a short time period. Furthermore, it has been recently demonstrated that semantic shift is a gradual process and not a sudden and distinctive phenomenon (Rosenfeld and Erk, 2018). Therefore, here we treat our task as a *word ranking problem*, where our aim is to rank the words based on their semantic shift. Importantly, this also enables us to validate the performance of our models in a more robust way as compared to treating the task as a classification problem, since in the latter case the precision score

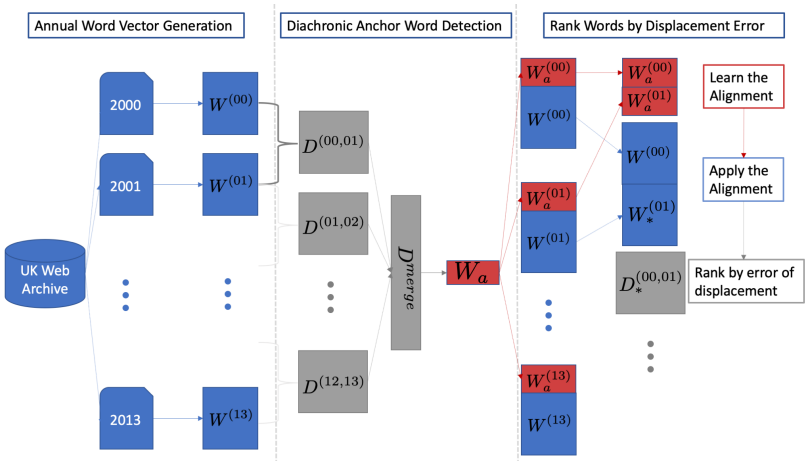

Figure 1: After constructing the word vectors on an annual basis, we learn their pairwise alignments of the resulting word vectors $\{W^{(t)}, W^{(t+1)}\}$. We rank the words based on their average displacement errors across all pairwise alignments in D^{merge} and select the k most stable words as our diachronic anchors W_a. Finally, we learn the alignment of $W_a^{(t+1)}$ based on $W_a^{(t)}$, and apply the same transformation to $W^{(t+1)}$ based on $W^{(t)}$. The words whose meaning has changed the most within $[t, t+1]$ are the ones with the largest displacement error in $D_*^{(t,t+1)}$.

of the positive (semantically shifted) class can be highly biased due to the small number of words belonging to it.

3.2 Our Approach

Figure 1 provides an overview of our approach for ranking the words based on their semantic shift levels. Given some word representations $\{W^{(t)}, W^{(t+1)}\}$ across two consecutive years (see section 4), our goal is to find an optimal way to align $W^{(t+1)}$ based on $W^{(t)}$, so that we can then compute the semantic shift level of a word by means of some distance metric. Typically, this alignment between $W^{(t+1)}$ and $W^{(t)}$ is performed on the complete vocabulary (Hamilton et al., 2016). This implies that the representation of words that have undergone semantic shift are still used as an input to the alignment algorithm, which can result into noisy pairwise alignments (Lubin et al., 2019).

To mitigate this issue, inspired by recent work in word translation (Conneau et al., 2017), we propose the use of a small number of *"anchor words"* to learn the optimal alignment between word representations at two consecutive time periods. Anchor words are defined as words whose lexical semantics remain static over two consecutive time periods. Similarly, *"diachronic anchor words"* correspond to those whose representations remain static across multiple and consecutive time intervals. The detection of these words can lead to more appropriate pairwise alignments of the word vectors, thus facilitating the task of finding semantically shifted words in a more robust fashion.

Anchor Words We formulate our approach on aligning the word vectors $\{W^{(t)}, W^{(t+1)}\}$ across consecutive time periods $[t, t+1]$ on the basis of the Orthogonal Procrustes problem (Schönemann, 1966). Besides past work on semantic change (Hamilton et al., 2016), this approach has been employed in related NLP tasks, such as word translation (Conneau et al., 2017; Ruder et al., 2018). In our case, it finds the optimal transformation of $W^{(t+1)}$ that best aligns it with $W^{(t)}$, by:

$$R = \underset{\Omega;\Omega^T\Omega=I}{\operatorname{argmin}} \left\| W^{(t)}\Omega - W^{(t+1)} \right\|_F. \quad (1)$$

The solution to Eq. 1 can be found via singular value decomposition: $R = UV^T$, where $U\Sigma V^T = SVD(W^{(t+1)}W^{(t)T})$. In our work, we ensure that $W^{(t+1)}$ and $W^{(t)}$ are centered at the origin and that $tr(W^{(t)}W^{(t)T}) = tr(W^{(t+1)}W^{(t+1)T}) = 1$. Finally, we transform $W^{(t+1)}$ as: $W_*^{(t+1)} = W^{(t+1)}R^T s$, where $s = \sum\Sigma$. We measure the displacement error matrix $D^{(t,t+1)}$ using the cosine distance over the resulting representations $\{W^{(t)}, W_*^{(t+1)}\}$. The k anchor words across $[t, t+1]$ correspond to the k words of $D^{(t,t+1)}$ with the lowest cosine distance (where one can vary the "stability" threshold of the anchor

words by varying k).

Diachronic Anchor Words The sets of the detected anchor words may vary between consecutive pairwise time intervals $\{[t, t+1], [t+1, t+2], ...\}$. This contrasts with our intuition of aligning the word vectors based on a few static (from a lexical semantic point of view) words. An intuitive way to accommodate this is to use words that are static throughout a longer period of time. Therefore, to detect "diachronic anchor words", we first perform all of the pairwise alignments and calculate the cosine distances of the words as before. We then concatenate these distances in a $|W|$-by-$|T|$ matrix D^{merge}. The diachronic anchor words correspond to the k words with the lowest average cosine distance in D^{merge}. In Figure 1, we denote their representations as W_a.

Semantic Change Detection We can now use a two-fold process to align the word vectors of two consecutive years $[t, t+1]$: first, we use Procrustes to learn the alignment of $W_a^{(t+1)}$ based on $W_a^{(t)}$, where $W_a^{(i)}$ corresponds to the vector representations of the diachronic anchor words at the time period i. Then, the learned transformation is applied to the representations of the complete vocabulary $W^{(t+1)}$, which are transformed into $W_*^{(t+1)}$. This way, we map the word representations at $t+1$ to the corresponding ones at t in a more robust way. Finally, we calculate the cosine distance matrix $D_*^{(t,t+1)}$ between the word representations in $W^{(t)}$ and $W_*^{(t+1)}$, where lower ranks indicate the index of a word with a higher level of semantic shift. The process is repeated for every pair of consecutive years, by keeping the same set of k diachronic anchor words for each alignment.

4 Data

We employ two datasets in our analysis: (a) the UK Web Domain Dataset 1996-2013 (JISC-UK) is used to learn word representations over different time periods (section 4.1); (b) the Oxford English Dictionary (OED) is used to refine our vocabulary and to build our ground truth – i.e., words that have changed their meaning over time (section 4.2).

4.1 JISC-UK Dataset

The UK Web Domain Dataset 1996-2013 (JISC-UK) contains textual information published in UK-based websites over the time period 1996-2013, thus facilitating the task of semantic change detection in a short-term and fine-grained temporal resolution (Basile and McGillivray, 2018).

Word Vectors Generation The dataset was processed based on previous work by Basile and McGillivray (2018), resulting in over 20TB of textual data. Instead of generating a single vector representation of a word across all years (e.g., by using Temporal Random Indexing (Basile and McGillivray, 2018)), we treated the concatenated content that was published within each year as a single (annual) document $D^{(t)}$, $t \in \{2000, ..., 2013\}$[1]. Following most of the past approaches on semantic change (Kim et al., 2014; Hamilton et al., 2016), we generated word representations by training $|T+1|$ `word2vec` models $m^{(t)}$ (one per year), using Skip-Gram with Negative Sampling and excluding all words appearing less than 1,000 times within a year. Each model was trained for five epochs, using a window size of five words. Finally, we represent every word in year t as a 100-dimensional vector $w_i^{(t)}$, and the resulting matrix of all words as $W^{(t)}$. The resulting vocabulary size per year is shown in Figure 2.

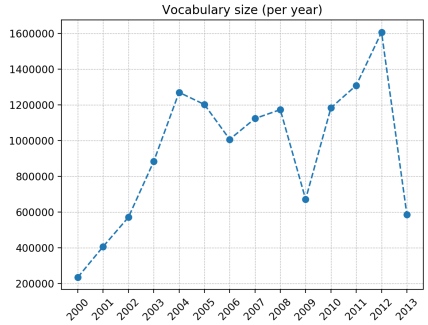

Figure 2: Vocabulary size per year (till May '13), excluding words appearing less than 1,000 times.

4.2 Oxford English Dictionary

The Oxford English Dictionary (OED) is one of the largest dictionaries and the most authoritative historical dictionary for the English language. It records over 250K lemmata along with their definitions, including the year in which each sense was first introduced in the language.

Ground Truth We consider the lemmata that are single words and with definitions whose first appearance in OED is recorded between 2001 and

[1] We excluded the years 1996-1999 owed to the data sparsity observed for these years.

2013 as our ground truth (218 words). Arguably, we expect there to be cases of words whose meaning has changed over time but are not recorded in the OED – i.e., the precision rate of our ground truth is not guaranteed to be 100%. However, this does not affect much our evaluation, since we are not treating our task as a classification problem, but are instead interested in ranking words with known semantic shift in an appropriate manner compared to more semantically stable words (i.e., we are interested in having high recall score of our ground truth, which is guaranteed by the OED).

4.3 Resulting Dataset

As opposed to early work studying the change in word frequency over time to detect semantic change (Michel et al., 2011), here we are interested in detecting words that have undergone semantic change based purely on their context. Therefore, to avoid any bias towards words that have appeared at a certain point in time (e.g., "facebook"), we focus strictly on the words that appear every year, yielding a vocabulary of 168,362 unique words. Finally, we filter out any word that does not appear in OED, due to the lack of ground truth for these words. The resulting dataset that is employed in our modelling is composed of **47,886** unique words that are present in OED and appear at least 1,000 times in every single year between 2000 and 2013, out of which **65** are marked by OED as words that have gained a new meaning after the year 2000[2].

4.4 Empirical Evidence of Semantic Change

Before presenting our experiments, it is important to get some insights on whether (a) semantic change actually occurs in such a limited time period and (b) that our ground-truth showcases this shift, in a qualitative manner.

We begin our analysis by leveraging Procrustes alignment in all possible year-to-year combinations and measure the sum of squared errors of each of the respective alignments. We try this approach on both (a) the intersected vocabulary (approximately 168K words) and (b) the resulting vocabulary by keeping only the words that are mapped to an OED entry (approximately 48K words, see section 4.3).

The results of this process are illustrated in Figure 3. We plot three heatmaps for each case,

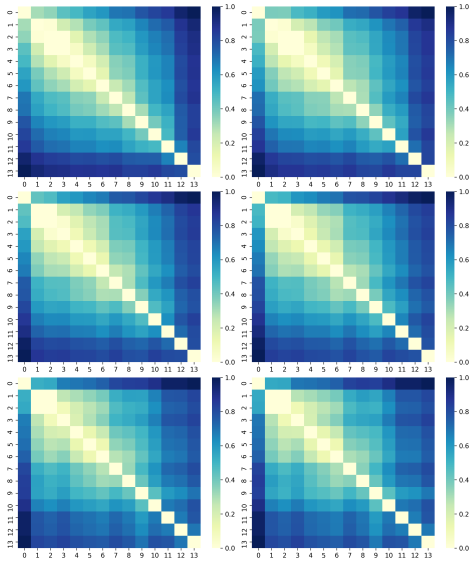

Figure 3: Normalised sum squared errors when aligning the word vectors across different years (2000-2013), using the complete vocabulary (left) and the its intersection with the OED dictionary (right), with different minimum frequency thresholds: 1K (top), 10K (middle) and 100K (bottom).

so that we see if there is an influence stemming from relatively rarely appearing words (when the threshold is set to 1,000). The results demonstrate that the further we move away from the diagonal, the higher the error becomes – note that this is picked up even though there is no notion of "time" in the alignments – indicating that there is a gradual/temporal shift in the meaning of the words, as captured by the context they appear in.

To further validate the notion of semantic change with respect to our ground truth, we take a closer look at the 65 semantically shifted words that are used in our experiments. The five closest neighbours (by means of cosine similarity) of eight of these words in the years 2000 and 2013 are shown in Figure 4, along with the shift level, measured for each word w and its neighbour n as $cos(w^{(13)}, n^{(13)}) - cos(w^{(00)}, n^{(00)})$. Figure 5 shows the temporal shift in the meaning of the same words from their top-100 neighbours in 2000, using a 3-year moving averaging filter. Both figures demonstrate that the semantic change of our ground-truth is captured through our representations. However, it is also demonstrated that semantic shift is a gradual process and its level may vary across different words. In what follows, we examine the extent to which we can capture this effect using our approach presented in section 3.2.

[2]The resulting dataset is available through: https://github.com/adtsakal/Semantic_Change

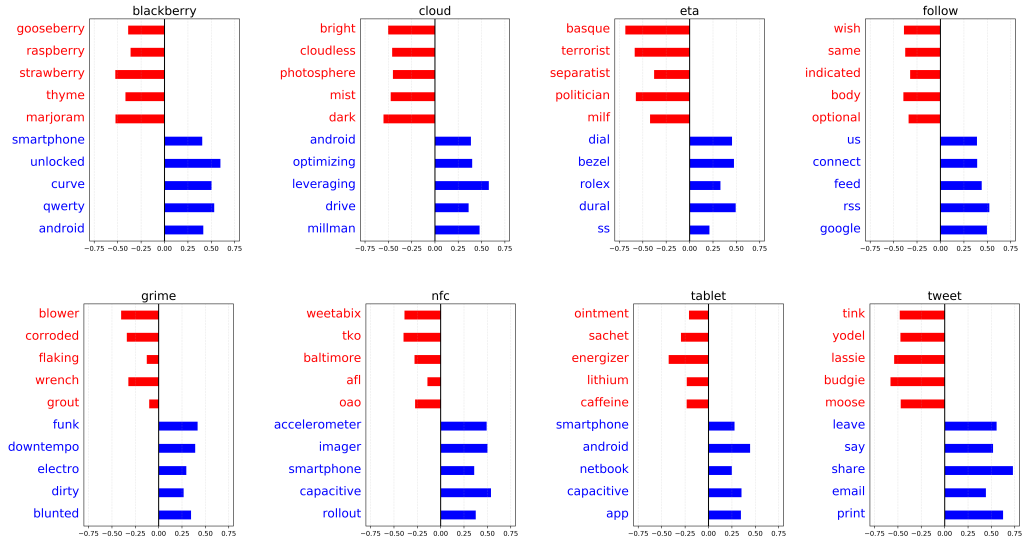

Figure 4: Closest neighbours of words that have undergone semantic shift, at two years (2000, 2013). The bar indicates the shift level of each word towards (away from) each of its neighbours in 2013 (2000).

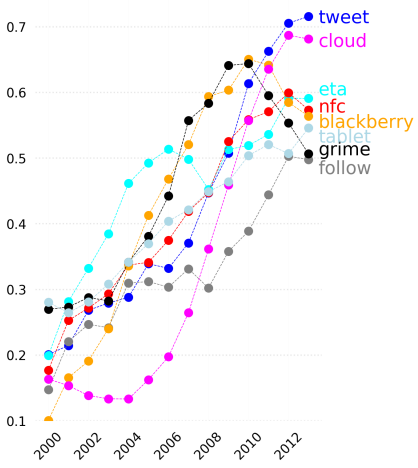

Figure 5: Cosine distance over time between four semantically shifted words (as marked by OED) and their top-100 neighbours in the year 2000.

5 Experiments

5.1 Task Formulation

Given the vector representations of all words across consecutive pairs of years (i.e., {[2000, 2001], ..., [2012, 2013]}), *our aim is to rank the words* based on their respective displacement errors that result after each pairwise alignment. Similarly to past work, we assume that the words corresponding to the highest displacement are those whose semantics has changed the most (Kim et al., 2014; Hamilton et al., 2016). The displacement of a word in a certain interval of a pair of years is calculated on the basis of the cosine distance between its resulting vectors on the first and the second year. Our task is performed on every pair of consecutive years separately.

5.2 Data Split

Experiment 1 We split our data into two sets: (a) in our training set we use most of our data to learn the alignment of the word representations across two different years, by ensuring that none of the 65 words denoted by the OED as words with altered meaning falls in this set (i.e., all of them are considered "static"); (b) we use the rest of our data for evaluation purposes (see next subsection), by ensuring that we include the 65 "changed" words in this set. We experiment with different percentage splits between training and evaluation sets (evaluation set size: [10%, ..., 50%]). This enables us to study the effect of the training set size and the number of diachronic anchor words (see 5.4 below) that are needed to detect semantic change effectively. Due to the small number of "changed" words in the evaluation set, for each percentage split, we perform 40 runs with random splits of the "static" words into the two sets.

Experiment 2 Here we use the complete set of word representations to learn the alignments across the different time intervals, disregarding the split into train/evaluation sets. This enables us to get clearer insights on the performance of the models under a complete setting and study the effect of diachronic anchor words in more detail.

5.3 Evaluation

We propose an alternative, rank-based metric, which can yield robust comparisons across different models, even with a relatively small number of labelled words. Given the final word rankings of an algorithm when applied on a certain pair of years, we denote *the average relative rank of a word whose meaning has changed* (as denoted by the OED) as μ-rank. The value of a single word for this metric lie within the $[0, 1]$ interval, with lower values indicating a better rank produced by the model. The μ-rank of a model is calculated for each of the 13 pairs of years independently and all the results are averaged across the 40 runs, yielding a vector of rank 13. Finally, we consider the average μ-rank score of this vector as our evaluation metric. For all of the models used in Experiment 1, the μ-rank is calculated based on the evaluation set.

5.4 Models

Baselines Our first vanilla approach ($\boldsymbol{PROCR_{100}}$) ranks the words by means of their respective displacement errors (i.e., cosine distance), by learning a single transformation across the whole dataset (Hamilton et al., 2016). For Experiment 1, we also include a second approach ($\boldsymbol{PROCR_{90}}$) which similarly learns the transformation based on the training set and then applies it to the evaluation set.

Our Models We employ two models based on the notion of anchor words: for a given pair of years, $\boldsymbol{PROCR_k}$ first learns an optimal alignment based on the full training set (similarly to $PROCR_{90}$) and then selects the k words with the lowest displacement error of this set to serve as "anchor" words, in order to learn a new alignment based strictly on them; this new transformation is then applied in the evaluation set to yield the final word rankings. This implies that the anchor words are not necessarily the same across all pairs of years. $\boldsymbol{PROCR_{kt}}$ operates in a similar fashion, albeit resolving this drawback through the use of diachronic anchor words: it first learns all of the alignments across the different pairs of years ({[2000, 2001], ..., [2012, 2013]}) and then it selects the k words with the lowest average displacement error across time; finally, it ranks the words in the evaluation set by learning a single transformation for every pair of years based strictly on these anchor words. For both of our models, we experiment with a varying value for k, measured as a % of the size of our training set in Experiment 1. For Experiment 2, we fix k to be the optimal number of words found in Experiment 1.

6 Results

Experiment 1 The results of our models and the baselines are presented in Figure 6. We provide one chart for each evaluation set percentage of the data that was used in our experiments, averaged over the 40 randomised splits we performed.

It becomes apparent that the anchor-based approaches perform clearly better (i.e., they have consistently lower average μ-rank) than those based on the alignment of all of the words – either of the training set, in $\boldsymbol{PROCR_{90}}$, or of both sets, in $\boldsymbol{PROCR_{100}}$. This is because the alignments of the former are based on the representations of words that are indeed stable over time. As we have empirically demonstrated in the previous section, semantic change is a gradual process; thus, aligning words whose representations are not stable across time results into noisy alignments that fail to capture the semantic change of words effectively.

The comparison between the anchor ($\boldsymbol{PROCR_k}$) and diachronic anchor ($\boldsymbol{PROCR_{kt}}$) approaches indicates that the latter performs consistently better. We find that using a very small number of anchor words (0.1% of the training set) yields much better results in almost all cases. Depending on the size of the evaluation set, this number of words ranges from 43 (in the case of 10%) down to 28 words (in the case of 40%). When we further increase the size of the evaluation set (thus decreasing the size of the training set) to 50% of our dataset, we find that using 1% (239) of the words in the training set as anchor words yields slightly better results than using 0.1%. This is because some of the anchor words are placed within the evaluation set, thus the alignment is learned based on weaker anchors, yielding poorer performance. Having a large training set to extract the (diachronic) anchor words from and learn the optimal alignments between their representations across different years is sufficient to overcome this issue.

Finally, while the proposed models outperform the standard practices found in related work, we observe that their performance is still relatively poor: a semantically shifted word is expected to be

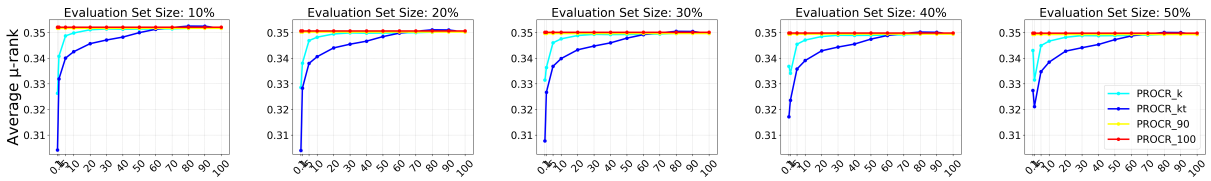

Figure 6: Average μ-rank in Experiment 1 across all runs, using different % of anchor words (x-axis).

ranked close to the top-30% of all of the competing words, with respect to its semantic shift level. This indicates that the task of semantic change detection is rather challenging. Incorporating the temporal dimension of the task is a promising direction for future research in this perspective.

Experiment 2 We present the results when we employ the full dataset to learn the alignments of the $\boldsymbol{PROCR_{100}}$, $\boldsymbol{PROCR_{k}}$ and $\boldsymbol{PROCR_{kt}}$ models. We fix the percentage of (diachronic) anchor words to be the 47 most stable words (i.e., the top-0.1%). The results are provided in Figure 7 in a per-year basis. $\boldsymbol{PROCR_{kt}}$ performs better on average μ-rank terms (29.48$\pm$3.67) against $\boldsymbol{PROCR_{k}}$ (32.68$\pm$4.93) and $\boldsymbol{PROCR_{100}}$ (35.08$\pm$4.71), demonstrating again the effectiveness of the alignment based on the diachronic anchor words.

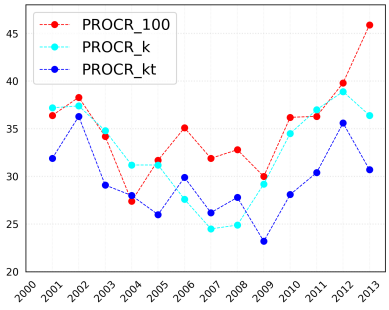

Figure 7: μ-rank of the three models on an annual basis in Experiment 2.

To shed light into the difference between the performance of models employing the anchor and the diachronic anchor words, we calculate the number of anchor words that belong to the set of the diachronic anchors, per year. On average, we find that only 16% (st.dev.: 5.9%) of the annually detected anchor words belong to the latter set. Throughout the pairwise alignments, there are overall 434 unique anchor words detected, from an overall possible of 611. This is owed to the "noisy" selection of anchor words. In Experiment 1, we have demonstrated that aligning the word vectors based on a very small number of anchors performs better. However, the accurate selection of such a small proportion of words can be rather challenging and can vary a lot over consecutive time intervals, due to the noisy nature of the word representations and the alignments themselves. By selecting diachronic anchor words, we are able to filter out this noise, thus yielding more accurate word alignments and tracking the semantic shift of words through time in a more robust way.

7 Conclusion and Future Work

We have introduced a new labelled dataset for semantic change detection. Approaching our task as a word ranking problem, we have proposed an approach to align word representations across different points in time on the basis of a few stable words across time. Through extensive experimentation, we have demonstrated that our approach yields better performance compared to current practices that are based on aligning word representations at different points in time.

An extension to our work is the incorporation of Generalised Procrustes Alignment (Gower, 1975). This will allow us to align the word representations across all years simultaneously and observe the trajectory of each word through time. Furthermore, in our exploratory analysis, we have qualitatively demonstrated that semantic change is a gradual process. Therefore, incorporating the temporal dimension of the task in our approach is a major direction for future work. In particular, we plan to incorporate temporal approaches that are well-suited for the task, such as temporal word clustering and change point detection. Finally, by making our resources publicly available, we hope to facilitate further research in the domain.

Acknowledgments

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the seed funding grant SF099.

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings via skip-gram filtering. *stat* 1050:27.

Pierpaolo Basile and Barbara McGillivray. 2018. Exploiting the Web for Semantic Change Detection. In *International Conference on Discovery Science*. Springer, pages 194–208.

Leonard Bloomfield. 1933. Language.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087* .

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7(2):121–157.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2018. Short-term meaning shift: an exploratory distributional analysis. *arXiv preprint arXiv:1809.03169* .

John C Gower. 1975. Generalized procrustes analysis. *Psychometrika* 40(1):33–51.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*. pages 67–71.

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)*. volume 1, pages 1489–1501.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE Press, pages 229–238.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. pages 61–65.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 625–635.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*. pages 1384–1397.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. 2019. Aligning Vector-spaces with Noisy Supervised Lexicons. *arXiv preprint arXiv:1903.10238* .

Jan Lukeš and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 65–71.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pages 259–263.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. pages 3111–3119.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pages 474–484.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A Discriminative Latent-Variable Model for Bilingual Lexicon Induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 458–468.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 1003–1011.

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

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6(12):e26752.

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

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering* 24(5):649–676.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the Stance of Social Media Users in a Sudden Vote: The Case of the Greek Referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, pages 367–376.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, pages 673–681.