# Exploiting the Web for Semantic Change Detection

Pierpaolo Basile[1] and Barbara McGillivray[2,3]

[1] Department of Computer Science, University of Bari Aldo Moro, Italy
pierpaolo.basile@uniba.it
[2] Modern and Medieval Languages, University of Cambridge, UK
[3] The Alan Turing Institute, London, UK
bmcgillivray@turing.ac.uk

**Abstract.** Detecting significant linguistic shifts in the meaning and usage of words has gained more attention over the last few years. Linguistic shifts are especially prevalent on the Internet, where words' meaning can change rapidly. In this work, we describe the construction of a large diachronic corpus that relies on the UK Web Archive and we propose a preliminary analysis of semantic change detection exploiting a particular technique called Temporal Random Indexing. Results of the evaluation are promising and give us important insights for further investigations.

**Keywords:** Semantic Change Detection · Diachronic Analysis of Language · Time Series.

## 1 Introduction

Languages can be studied from two different and complementary viewpoints: the diachronic perspective considers the evolution of a language over time, while the synchronic perspective describes the language rules at a specific point of time, without taking its history into account [8]. During the last decade, the surge in available data spanning different epochs has inspired a new analysis of cultural, social, and linguistic phenomena from a temporal perspective. Language is dynamic and evolves, it varies to reflect the shift in topics we talk about, which in turn follow cultural changes. So far, the automatic analysis of language has largely been based on datasets that represented a snapshot of a given domain or time period (*synchronic approach*). However, since the rise of big data, which has made large corpora of data spanning several periods of time available, large-scale diachronic analysis of language has emerged as a new approach to study linguistic and cultural trends over time by analysing these new sources of information. One of the largest sources of information is the Web, which has been exploited to build corpora used in linguistics or in Natural Language Processing (NLP) tasks. Generally, these corpora are built using a synchronic approach without taking into account temporal information.

In this paper, we propose to analyze the Web using a *diachronic approach* by relying on the UK Web Archive project [15]. The goal of this project is to analyse the change in language over time as reflected in the textual content of UK websites. We focus on one specific kind of language change, namely semantic change, aiming to develop a computational system that is able to detect which words have changed meaning over the period of time covered by the corpus of UK websites.

Semantic change is a very common phenomenon in language. Over time, words can acquire new meanings or lose existing ones. For example, the original meaning of the verb *tweet*, according to the Oxford English Dictionary (OED), is transitive, defined as follows:

```
Of a bird: to communicate (something) with a brief
high-pitched sound or call, or a series of such sounds.
```

According to the OED, this meaning was first recorded in writing in 1851. On the other hand, the OED assigns the first written usage of the related intransitive meaning to 1856:

```
Of a bird: to make a brief high-pitched sound or call, or
a series of such sounds. Also in extended use.
```

The OED also lists two additional senses, which are much more recent. The transitive one is defined as follows:

```
To post (a message, image, link, etc.) on the social
networking service Twitter. Also: to post a message
to (a particular person, organization, etc.).
```

This meaning was first recorded in 2006. The intransitive one is defined as:

```
To make a posting on Twitter. Also: to use Twitter
regularly or habitually.
```

and was first recorded in 2007.

Semantic change detection systems allow for large-scale analyses that identify cultural and social trends. For example, when the contexts of the word *sleep* are compared between 1960s and 1990s, it has been shown through distributional semantics models that this word acquired more negative connotations linked to sleep disorders [12]. Moreover, such systems have a range of applications in NLP. For example, they can improve sentiment analysis tools because they can identify positive or negative content expressed via newly emerged meanings, such as the positive slang sense of *sick* meaning “awesome”.

The use of the Web as a source of data for diachronic semantic analysis poses an important challenge that we aim to tackle in this paper: the massive size of the dataset requires efficient computational approaches which are able to scale up to process terabytes of data. In this scenario, Distributional Semantic Models (DSMs) represent a promising solution. DSMs are able to represent words as points in a geometric space, generally called WordSpace [23, 22] by simply analysing how words are used in a corpus. However, a WordSpace represents a snapshot of a specific corpus and it does not take into account temporal information. For this reason, we rely on a particular method, called Temporal Random Indexing (TRI), that enables the analysis of the time evolution of the meaning of a word [4, 16]. TRI is able to efficiently build WordSpaces taking into account temporal information. We exploit this methodology in order to build geometrical spaces of word meanings that span over several periods of time. The TRI framework provides all the necessary tools to build WordSpaces over different time periods and perform such temporal linguistic analysis. The system has been tested on several domains, such as a collection of Italian books, English scientific papers [3], the Italian version of the Google N-gram dataset [2], and Twitter [16].

The paper is structured as follows: Section 2 provides details about our methodology, while Section 3 describes the dataset that we have developed and the results of a preliminary evaluation. Related work is provided in Section 4, and Section 5 reports final remarks and future work.

## 2 Method

This section provides details about the methodology adopted during our research work. In particular, we build a diachronic corpus using data coming from the Web. Relying on this corpus, we build a semantic distributional model that takes into account temporal information. The last step is to build time series in order to track how the meaning of a word change over time. These time series are created by exploiting information extracted from the distributional semantic models. In the following sub-sections we provide details about each of the aforementioned steps.

### 2.1 Corpus creation

The first step is to create a diachronic corpus starting from data coming from the web. The web collection under consideration is the JISC UK Web Domain Dataset (1996-2013) [15] which collects resources from the Internet Archive (IA) that were hosted on domains ending in .uk, and those that are required in order to render .uk pages.

The JISC dataset is composed of two parts: 1) the first part contains resources from 1996 to 2010 for a total size of 32TB; 2) the second one contains resources from 2011-2013 for a total size of 30TB. The JISC dataset cannot be made generally available, but can be used to generate secondary datasets. For that reason we provide the corpus in the form of co-occurrence matrices extracted from it. The dataset contains resources crawled by the IA Web Group for different archiving partners, the Web Wide crawls and other miscellaneous crawls run by IA, as well as data donations from Alexa and other companies or institutions. So it is impossible to know all the crawling configuration used by the different partners. However the dataset contains not only HTML pages and textual resources but also video, images and other types of files.

The first step of the corpus creation consists in filtering the JISC dataset in order to extract only the textual resources. For this purpose, we extract the text from textual resources (e.g. TXT files) and parse HTML pages in order to extract their textual content. We adopt the jsoup library[4] for parsing HTML pages.

The original dataset stores data in the ARC and WARC formats, which are standard formats used by the Internet Archive project for storing data crawled from the web

[4] https://jsoup.org/

as sequences of content blocks. The WARC format is an enhancement of ARC for supporting metadata, detect duplicate events and more. We process ARC and WARC archives in order to extract the textual content and store data in the WET format. WET is a standard format for storing plain text extracted from in ARC/WARC archives. We transform the original dataset in the standard WET format which contains only textual resources. The output of this process provides about 5TB of WET archives.

The second step consists in tokenizing the WET archives in order to produce a tokenized version of the textual content. We exploit the StandardAnalyzer.[5] provided by the Apache Lucene API[6] This analyzer provides also a standard list of English stop words. The size of the tokenized corpus is approximately 3TB.

In the third step, we create co-occurrence matrices, which store co-occurrences information for each word token. In order to track temporal information, we build a co-occurrence matrix for each year from 1996 to 2013. Each matrix is stored in a compressed text format, one row per token. Each row reports the target token and the list of tokens co-occurring with it. An example for the word *linux* is reported in Figure 1, which shows that the token *swapping* co-occurs 4 times with *linux*, the word *google* 173 times, and so on. We extract co-occurrences taking into account a window of five words to the left and to the right of the target word. For the construction of co-occurrence matrices, we exploit only words that occur at least 4,500 times in the dataset. We do not apply any text processing step such us lemmatization or stemming for two reasons: 1) the idea is to build a language independent tool and 2) in this first evaluation we want to reduce the number of parameters and focus our attention on the change point detection strategy. Finally, we obtain a vocabulary of about one million words and the total size of compressed matrices is about 818GB.

```
linux   swapping   4   google   173   xp   454   manufacturer
237   job   64   install   255   security   137   cgi   47
operating   705   host   69   performance   44   sharing
56...
```

**Fig. 1.** Co-occurrence matrix

The whole process is described in Figure 2: WARC/ARC archives are converted into WET files in order to extract the text and they are tokenized; the tokenized text is exploited by the *Matrix Builder* for building the co-occurrence matrices; matrices are the input for *TRI* that performs Temporal Random Indexing and provides a WordSpace for each time period; finally WordSpaces are used to build time series. The last part of the chart sketches the process used to detect semantic changepoints (see Subsection 2.2) and the evaluation step described in Section 3.

[5] https://lucene.apache.org/core/7_3_1/core/index.html
[6] https://lucene.apache.org/core/

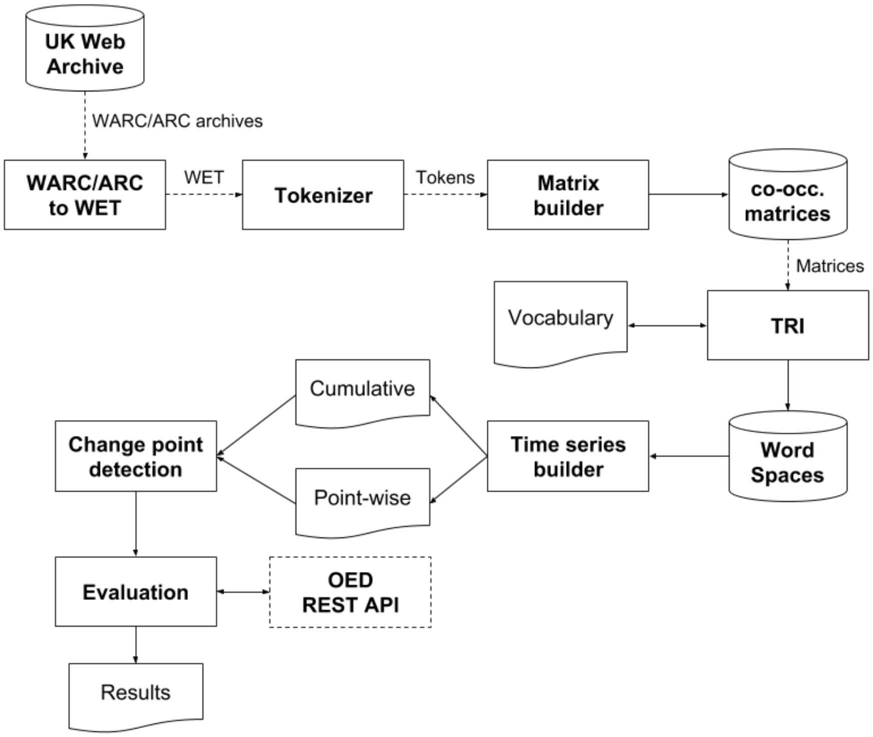

**Fig. 2.** Flowchart of the whole semantic change detection process.

### 2.2 Semantic change detection

Our method for semantic change detection relies on a previous model based on Temporal Random Indexing (TRI) [3, 4]. In particular, we further develop the TRI approach in three directions: 1) we improve the system in order to manage very large datasets, such as the JISC UK Web Domain Dataset; 2) we introduce a new way to weight terms in order to reduce the impact of very frequent terms; 3) we introduce new methods for detecting semantic shift from time series analysis techniques.

The idea behind TRI is to build different WordSpaces for each time period under investigation. The peculiarity of TRI is that word vectors over different time periods are directly comparable because they are built using the same random vectors. TRI works as follows:

1. Given a corpus $C$ of documents and a vocabulary $V$ of terms[7] extracted from $C$, the method assigns a random vector $r_i$ to each term $t_i \in V$. A random vector is a vector that has values in {-1, 0, 1} and is sparse with few non-zero elements randomly distributed along its dimensions. The sets of random vectors assigned to all terms in $V$ are near-orthogonal;

[7] $V$ contains the terms that we want to analyse, typically, the most $n$ frequent terms.

2. The corpus $C$ is split into different time periods $T_k$ using temporal information, for example the year of publication;
3. For each period $T_k$, a WordSpace $WS_k$ is built. All the terms of $V$ occurring in $T_k$ are represented by a semantic vector. The semantic vector $sv_i^k$ for the $i$-th term in $T_k$ is built as the sum of all the random vectors of the terms co-occurring with $t_i$ in $T_k$. When computing the sum, we apply some weighting to the random vector. In our case, to reduce the impact of very frequent terms, we use the following weight: $\sqrt{\frac{th \times C_k}{\#t_i^k}}$, where $C_k$ is the total number of occurrences in $T_k$ and $\#t_i^k$ is the occurrences of the term $t_i$ in $T_k$. The parameter $th$ is set to 0.001.

This way, the semantic vectors across all time periods are comparable since they are the sum of the same random vectors.

In order to track the words' meaning change over time, for each term $t_i$ we build a time series $\Gamma(t_i)$. A time series is a sequence of values, one value for each time period, and it indicates the semantic shift of that term in the given period. We adopt several strategies for building the time series. The first strategy is based on term log-frequency; each value in the series is defined as: $\Gamma_k(t_i) = log(\frac{\#t_i^k}{C_k})$.

In order to exploit the ability of our methods in computing vectors similarity over time periods, we define two strategies for building the time series:

**point-wise:** $\Gamma_k(t_i)$ is defined as the cosine similarity between the semantic vector of $t_i$ in the time period $k$, $sv_i^k$, and the semantic vector of $t_i$ in the previous time period, $sv_i^{k-1}$. This way, we aim to capture semantic change between two time periods;

**cumulative:** we build a cumulative vector $sv_i^{C_{k-1}} = \sum_{j=0}^{k-1} sv_i^j$ and compute the cosine similarity of this cumulative vector and the vector $sv_i^k$. The idea behind the cumulative approach is that the semantics of a word at point $k-1$ depends on the semantics of the word in all the previous time periods. The cumulative vector is the semantic composition of all the previous word vectors, the composition is performed through the vector sum [20].

Given a time series, we need a method for finding significant changepoints in the series, which we interpret as indications that semantic change has taken place. We adopt three strategies:

1. the *Mean shift model* [26], proposed in [17], defines a mean shift of a general time series $\Gamma$ pivoted at time period $j$ as:

$$K(\Gamma) = \frac{1}{l-j} \sum_{k=j+1}^{l} \Gamma_k - \frac{1}{j} \sum_{k=1}^{j} \Gamma_k \tag{1}$$

In order to determine if a mean shift is relevant at time $j$, we adopt a bootstrapping [10] approach, under the null hypothesis that there is no change in the mean. In particular, a confidence level is computed by constructing $B$ bootstrap samples by permuting $\Gamma(t_i)$. Finally, we estimate changepoints by considering the time points with a confidence value above a predefined threshold;

2. the *valley model*, in which any point $j$ that has a value lower than the previous point $j - 1$ in the time series is considered a changepoint. The idea is that if we observe a decrease in the similarity between the semantic vector of a word at a given point in time and the semantic vector of the same word in the previous time point, then this indicates that the word's semantics is changing;
3. the *variance model*, in which the difference between the value in the time series at a point $j$ and the value at the point $j - 1$ is compared with the variance of the time series; when the difference is higher than one, two or four times the variance, the point is considered a changepoint.

### 2.3 System output and neighborhood analysis

The system's output consists of lists of candidate words which are predicted to have undergone semantic change, together with the year in which this change is predicted to have happened. In addition, for each candidate, we can extract its corpus neighbours, defined as the top $n$ words whose semantic vectors have the highest cosine similarity with the vector of the candidate word.

To take an example, our system considered *blackberry* as a candidate for semantic change, with three changepoints, in the years 1998, 2007, and 2009. The original sense of *blackberry* refers to the *"edible berry-like fruit of the bramble, Rubus fruticosus"*, the *"The trailing plant Rubus fruticosus"*, and *"Any of various other dark-coloured edible berries"*, according to the OED. However, a more recent sense emerged in 1999, defined in the OED as *"A proprietary name for: a type of pager or smartphone capable of sending and receiving email messages"*.

If we look at the top 20 neighbours of *blackberry* in 1999 extracted by TRI from the UK Web Archive JISC dataset 1996-2013 corpus, we see that the majority of them are words related to original sense (highlighted in bold face the list below), either as collocates (like *pie*) or as distributionally similar nouns to *blackberry* (like *strawberry*):

**cherry**, **berries**, **strawberry**, **blossom**, **pie**, **blueberry**, **blackcurrant**, brierley, **pudding**, beacon, **red**, **raspberry**, hill, lion, mill, green, **chestnut**, brick, **ripe**, **scent**

On the other hand, the top 20 corpus neighbours of *blackberry* in 2003 include some words from the domain of mobile phones, highlighted in bold face below[8]:

blueberry, plum, **phones**, **cellphones**, **handsets**, loganberry, ripe, strawberry, **devices**, orange, **phone**, currant, gooseberry, **gprs**, wings, blackcurrant, damson, **bluetooth**, berries, blackberries

By 2004, the majority of corpus neighbours of *blackberry* involve words related to mobile phones, indicating that this has become the predominant sense of the word in the corpus, as shown by the following list of top 20 neighbours in 2004:

[8] We did not highlight *orange* in the list because in this context it could refer to the fruit, in which case it would be related to the fruit sense of *blackberry*, or to the mobile phone company, in which case it would be related to the cellphone sense of *blackberry*.

handspring, **handsets**, **tmobile**, **justphones**, nec, **handset**, **payg**, **tarriffs**, **lg**, **cellphones**, **pickamobile**, **phonesnokia**, **prepay**, **sim**, **tariffs**, **phones**, **phoneid**, **findaphone**, **mobilechooser**, **unlock**

The system's output lists contain several thousand candidates, a set which is too large to assess by hand. Therefore, we devised a novel automatic evaluation framework, outlined in the next section.

## 3 Evaluation

There is no general framework for evaluating the accuracy of semantic change detection systems. Previous work has evaluated semantic change systems either indirectly via their performance on related tasks (e.g. [11]), or via a small-scale qualitative analysis (e.g. [13]). In order to measure how well our system achieves the intended aim to identify words that have changed their meaning over the time covered by the UK Web Archive JISC dataset 1996-2013, we developed a novel evaluation framework. We evaluated our semantic change system and a baseline system against a dictionary-based gold standard. In the baseline system, we used a time series consisting of the frequency counts of each word form in the corpus. The evaluation of this baseline was aimed to detect any contribution given by the cosine similarity scores and TRI in our system.

We used the data from the Oxford English dictionary (OED) API as gold standard. The OED contains a diachronic record of the semantics of the words in the English lexicon. Each entry corresponds to a lemma and part-of-speech pair, and contains the list of its senses, each with a definition, the year when each sense was first recorded in writing, a corresponding quotation, following optionally more dated quotations which illustrate the use of the word with that sense at different points in time.

We performed the evaluation of each system in two steps. First, we calculated the accuracy of the semantic changepoint detection component, with the aim to measure how well the system detected semantic change candidates at the correct point in time. For each semantic change candidate outputted by each system, we checked that it appeared in the OED with a first usage dated from 1995 or later[9]. If this was not the case, we excluded the candidate word and the changepoint year from the analysis, as we were not able to assess whether the word changed meaning in the time span under consideration. We also only considered words that had a frequency of at least 100 in the corpus. We compared the changepoint year of semantic change according to our system with the year when the sense was first recorded according to the OED. The candidate and its changepoint were considered correct if the changepoint year was no earlier than the year when it was first recorded according to the OED. For example, the OED records the first usage of the verb *follow* with the transitive meaning of *"To track the activities or postings of (a person, group, etc.) by subscribing to their account on a social media*

[9] As the earliest texts in the corpus date from 1996, we allowed for a one-year buffer between this date and the date of first usage according to the OED, under the assumption that a sense first recorded in the OED in 1995 could be recorded with sufficient evidence in our corpus at least one year later.

*website or application.”*, and dates it from 2007. Our system suggested *follow* as a candidate for semantic change, with a changepoint in 2009. According to our evaluation approach, this counted as a correct candidate.

The results of the first evaluation step are summarized in Table 1. Semantic change detection is a very difficult task, especially when measured against a highly-curated resource like the OED, which relies on an evidence basis that is much broader in scope compared to the UK Web Archive. Therefore, it is not surprising that the precision scores are low. Of the several tens of thousands candidates outputted by our system or the baseline, only less than 400 were correct, in all configurations of the parameters. The precision scores range between 0.003 and 0.005. Given that the number of words in the gold standard is 462, the recall scores range between 0.104 and 0.849, with the highest score being associated to the point-wise and cumulative time series and the valley model for changepoint detection. It is important to note that methods reporting the highest recall (cumulative/valley and point-wise/valley) provide a high number of candidates (about 77,515) but these represent only the 7,7% of the whole dictionary exploited by our system (about one million). Overall, we can say that the valley model for changepoint detection yields the highest recall scores and outperforms the mean shift model and the variance model, and that the system with cumulative and pointwise time series outperforms the system with frequency-based time series (baseline). We are not able to provide a comparison with methods based on word embeddings due to the difficult to scale-up these approaches on our large corpus. We plan to perform this comparison as future work.

| System | changepoint | # correct | candidates | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| Baseline | Mean shift | 76 | 14,176 | 0.005 | 0.165 | 0.010 |
| Baseline | Valley | 378 | 77,493 | 0.005 | 0.818 | 0.010 |
| Baseline | Variance 1 | 0 | 145 | 0 | 0 | 0 |
| Baseline | Variance 2 | 0 | 52 | 0 | 0 | 0 |
| Cumulative | Mean shift | 48 | 15,266 | 0.003 | 0.104 | 0.006 |
| **Cumulative** | **Valley** | 392 | 77,515 | 0.005 | **0.848** | 0.010 |
| Cumulative | Variance 1 | 165 | 47,389 | 0.003 | 0.357 | 0.007 |
| Cumulative | Variance 2 | 56 | 14,452 | 0.004 | 0.121 | 0.008 |
| Point-wise | Mean shift | 74 | 23,855 | 0.003 | 0.161 | 0.006 |
| **Point-wise** | **Valley** | 392 | 77,515 | 0.005 | **0.848** | 0.010 |
| Point-wise | Variance 1 | 382 | 76,061 | 0.005 | 0.827 | 0.010 |
| Point-wise | Variance 2 | 340 | 69,492 | 0.005 | 0.736 | 0.010 |

**Table 1.** Summary of evaluation metrics of our systems and the baseline against the gold standard (OED). The first column details the time series construction type; the second column details the changepoint detection approach. The variance approach is followed by a numeric parameter: ‘Variance 1’ means that the changepoint is identified when the difference between the value in the time series at a point $j$ and the value at the point $j - 1$ is higher than the variance of the time series; ‘Variance 2’ means that the changepoint is identified when the difference between the value in the time series at a point $j$ and the value at the point $j - 1$ is higher than twice the variance of the time series.

For the second evaluation step, we focussed on the candidates that were considered correct according to the method explained above. For those candidates, we measured the accuracy of the output from the point of view of their semantics. In other words, we checked that the new meanings of the correct candidate words identified by the system corresponded to the new meanings as recorded in the gold standard. For each semantic change candidate word (and corresponding changepoint year) which was considered correct according to the approach illustrated above, we assessed how closely the new meaning of the candidate matched the senses in the OED first recorded after 1995. We measured this by collecting two sets of words. For the first set, we approximated the semantics of the new meaning as detected by the system with the 100 closest corpus neighbours to the candidate word, measuring proximity between words with the cosine distance. For the second set, we approximated the semantics of the OED senses with a bag-of-words approach. We pre-processed all words appearing in the definition and quotation text of each OED sense by stemming and lower-casing them. Then, we compared the two sets by calculating their Jaccard index, defined as the ratio between the number of elements in the intersection between the two sets, and the number of elements in the union of the two sets. Finally, we extracted the rank of each sense by according to the Jaccard index with the corpus neighbours (in decreasing order), and reported the rank of the correct candidate as an evaluation measure. For this evaluation, we focussed on the best-performing models according to the recall measure, as precision scores were low in all cases. These models involved collecting the time series with the pointwise and cumulative methods, and calculating the changepoint with the valley method and led to the highest recall score of 0.848.

Let us take the example of *mobile*, which the system predicted changed its meaning in 2000. *Mobile* has two post-1995 senses in the OED, the first is first recorded in 1998, and the second is first recorded in 1999. Their definitions from the OED are, respectively:

1. A person's mobile phone number; cf. mobile phone number n.
2. As a mass noun. Mobile phone technology, networks, etc., esp. considered as a means to access the Internet; the Internet as accessed from mobile phones, tablet computers, and other portable wireless devices. Frequently with on, over, via, etc.

The top 20 corpus neighbours for *mobile* in 2000 include the words *phones, phone, connected, devices*, which are shared with the OED definition and quotation of the second sense.

Table 2 shows the results of the neighbourhood-based evaluation on the best performing models according to recall scores. Although the Jaccard indices between the corpus neighbours of the candidates and the bag-of-words from the OED definition and quotation texts are usually very low, with an average of only 0.008, when we matched the semantics of the candidates (as measured by their top 100 corpus neighbours) with the OED senses first recorded after 1995, we found that the OED senses corresponding to the model's candidates (i.e. those OED senses whose first usage was no later than the candidates' changepoints) tended to be ranked first. This indicates that the models are accurate not only at spotting the correct changepoint for a word, but also its new semantic features.

| System | changepoint | Av. rank | Av. OED senses | Av. rank (> 1 sense) | # OED senses (> 1) |
|---|---|---|---|---|---|
| Cumulative | Valley | 1.206 | 1.336 | 1.811 | 2.324 |
| Point-wise | Valley | 1.206 | 1.332 | 1.799 | 2.290 |

**Table 2.** Results of the neighbourhood-based evaluation on the two models with highest recall scores. The third column shows the average rank of the matching OED senses of the candidates. The fourth column shows the average number of OED senses included in the ranking. The fifth column shows the average rank of the matching OED senses excluding the cases in which there is only one OED sense for the candidates. The last column shows the average number of OED senses included in the ranking, excluding the cases in which there is only one OED sense for the candidates. The ranking is based on the Jaccard index between the corpus neighbours of the candidate and the bag-of-words of the OED definition and quotation text.

In conclusion, analyzing the results we can notice that both cumulative and point-wise methods are able to overcome the baseline even though generally the precision is low due to the task difficulty. Evaluating semantic shift detection approaches is an open challenge, and researchers often rely on self-created test sets, or even simply manually inspecting the results. Moreover, our approach is able to correctly identify the semantics of the change according to the definition in the dictionary. We believe that this is the first work that tries to systematically analyze the semantic aspect of the changepoint.

## 4 Related work

Over the past decade, semantic change detection has been an emerging research area within NLP, and a variety of different approaches have been developed. Recent surveys on the current state of the art in this field have also been produced [24, 18].

A significant portion of the research in this area has focused on detecting semantic change in diachronic corpora spanning over several centuries [12, 27, 14, 28, 9, 11]. One of the most commonly used corpora is the multilingual Google Books N-gram Corpus [19], which covers the last five centuries and contains the N-grams from texts of over 6% of books ever published. Other researchers have used the 1800-1999 portion of this dataset, which consists of $8.5 * 10^{11}$ tokens [13].

A smaller set of previous studies focus on the more difficult task of detecting semantic change over a shorter time period, and use corpora which cover relatively short spans. Examples include a corpus consisting of articles from the New York Times published between 1990 and 2016 [29], a corpus based on the issues of the Chinese newspaper “People’s Daily” from 1946 to 2004 [25], the British National Corpus (100 million words, 1990s) [7], and data from the French newspaper “Le Monde” between 1997 and 2007 [6].

Concerning the methods employed, previous work includes a range of methods, from neural models to Bayesian learning [11] to various algorithms for dynamic topic modeling [5]. A significant part of the literature employ methods based on word embeddings [9, 13]. Very recently, dynamic embeddings have been shown as an improvement over using classical static embeddings for the task of semantic change detection [1, 21]. In this method, embedding vectors are inferred for each time period and a joint model is trained over all intervals, while simultaneously allowing word and context vectors to drift.

All previous works based on word embeddings have in common the fact that they build a different semantic space for each period taken into consideration; this approach does not guarantee that each dimension bears the same semantics in different spaces [16], especially when embedding techniques are employed. In order to overcome this limitation, Jurgens and Stevens [16] introduced Temporal Random Indexing technique as a means to discover semantic changes associated to different events in a blog stream. Our methodology relies on the technique introduced by Jurgens and Stevens, but with a different aim. While Jurgens and Stevens exploit TRI for the specific task of event detection, we setup a framework for semantic change detection relying on previous studies where TRI was applied on collection of Italian books, English scientific papers [3] and the Italian version of the Google N-gram dataset [2]. Moreover, it is important to stress that word embeddings techniques are based on word/context prediction that requires a learning step. On the other hand, TRI is based on counting words in context that is less computationally expensive and allows to scale up the method on a large Web collection.

## 5 Conclusion

In this work, we proposed several methods based on Temporal Random Indexing (TRI) for detecting semantic changepoints in the Web. We built a diachronic corpus exploiting the JISC UK Web Archive Dataset (1996-2013) which collects resources from the Internet Archive (IA) that were hosted on domains ending in .uk. We extracted about 5TB of textual data and we performed a preliminary evaluation using the Oxford English Dictionary (OED) as the gold standard. Results show that methods based on TRI are able to overcome baselines based on word occurrences, however, we obtain low precision due to a large number of detected changepoints. Moreover, for the first time, we propose a systematical approach for evaluating the semantics of detected changepoints by using both the neighborhood and the word meaning definition extracted from the OED. The precision of our model is low, which can be explained by several factors. First, the evaluation was based on an external resource, the OED, which relies on different data sources compared to web pages. This means that a semantic change recorded by our system is likely not to be necessarily reflected in the OED. Second, the task of semantic change detection is very hard, and our contribution is the first one to provide an evaluation based on a dictionary, so low precision values are not surprising. On the other hand, recall reaches a maximum value of 84%, which we consider an encouraging result. Overall, the results we report show that our approach is not only able to detect the correct time period, but also it is able to capture the correct semantics associated with the changepoint. As future work we plan to extend our analysis to the whole corpus and we want to investigate other time series approaches for reducing the number of detected changepoints with the aim of increasing the precision.

## Acknowledgments

This research was undertaken with the support of the Alan Turing Institute (EPSRC Grant Number EP/N510129/1). The access to the Oxford English Dictionary API was provided by Oxford University Press via a licence for non-commercial research.

## References

1. Bamler, R., Mandt, S.: Dynamic word embeddings. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 380–389. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017), `http://proceedings.mlr.press/v70/bamler17a.html`
2. Basile, P., Caputo, A., Luisi, R., Semeraro, G.: Diachronic analysis of the italian language exploiting google ngram (2016)
3. Basile, P., Caputo, A., Semeraro, G.: Analysing word meaning over time by exploiting temporal random indexing. In: Basili, R., Lenci, A., Magnini, B. (eds.) First Italian Conference on Computational Linguistics CLiC-it 2014. Pisa University Press (2014)
4. Basile, P., Caputo, A., Semeraro, G.: Temporal random indexing: A system for analysing word meaning over time. Italian Journal of Computational Linguistics **1**(1), 55–68 (12 2015)
5. Blei, D.M., Lafferty, J.D.: Dynamic topic models. ICML pp. 113–120 (2006)
6. Boussidan, A., Ploux, S.: Using Topic Salience and Connotational Drifts to Detect Candidates to Semantic Change. In: Proceeding IWCS '11 Proceedings of the Ninth International Conference on Computational Semantics. pp. 315–319 (2011)
7. Cook, P., Stevenson, S.: Automatically identifying changes in the semantic orientation of words. In: Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta (2010)
8. De Saussure, F.: Course in general linguistics. La Salle, Illinois: Open Court. (1983)
9. Dubossarsky, H., Weinshall, D., Grossman, E.: 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. pp. 1136–1145 (2017)
10. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. Chapman and Hall/CRC (1994)
11. Frermann, L., Lapata, M.: A bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics **4**, 31–45 (2016)
12. Gulordava, K., Baroni, M.: A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop. pp. 67–71 (2011), `http://clic.cimec.unitn.it/marco/publications/gems-11/gulordava-baroni-gems-2011.pdf`
13. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096 (2016)
14. Jatowt, A., Duh, K.: A framework for analyzing semantic change of words across time. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries pp. 229–238 (2014). https://doi.org/10.1109/JCDL.2014.6970173
15. JISC, the Internet Archive: Jisc uk web domain dataset (1996-2013) (2013). https://doi.org/https://doi.org/10.5259/ukwa.ds.2/1
16. Jurgens, D., Stevens, K.: Event Detection in Blogs using Temporal Random Indexing. In: Proceedings of the Workshop on Events in Emerging Text Types. pp. 9–16. Association for Computational Linguistics (2009)

17. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web. pp. 625–635. ACM (2015)
18. Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: a survey (2018), http://arxiv.org/abs/1806.03537
19. Lin, Y., Michel, J.B., Aiden, E.L., Orwant, J., Brockman, W., Petrov, S.: Syntactic annotations for the google books ngram corpus. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 8-14 July 2012. pp. 169–174. Association for Computational Linguistics (2012)
20. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive Science **34**(8), 1388–1429 (2010)
21. Rudolph, M., Blei, D.: Dynamic embeddings for language evolution. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web (2018)
22. Sahlgren, M.: The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces (2006)
23. Schiitze, H.: Word space. Advances in neural information processing systems **5**, 895–902 (1993)
24. Tang, X.: A State-of-the-Art of Semantic Change Computation. arXiv preprint arXiv:1801.09872 (Cl), 2–37 (2018)
25. Tang, X., Qu, W., Chen, X.: Semantic change computation: A successive approach. World Wide Web - Internet & Web Information Systems **19**(3), 375–415 (2016). https://doi.org/10.1007/s11280-014-0316-y
26. Taylor, W.A.: Change-point analysis: a powerful new tool for detecting changes. Taylor Enterprises, Inc. (2000)
27. Wijaya, D.T., Yeniterzi, R.: Understanding semantic change of words over centuries. Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web - DETECT '11 p. 35 (2011). https://doi.org/10.1145/2064448.2064475, http://dl.acm.org/citation.cfm?doid=2064448.2064475
28. Xu, Y., Kemp, C.: A Computational Evaluation of Two Laws of Semantic Change. Proceedings of CogSci 2015 pp. 1–6 (2015)
29. Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H.: Dynamic Word Embeddings for Evolving Semantic Discovery. Tech. rep. (2017). https://doi.org/10.1145/3159652.3159703, http://arxiv.org/abs/1703.00607{\%}0Ahttp://dx.doi.org/10.1145/3159652.3159703