Diachronic word embeddings and semantic shifts: a survey

Andrey Kutuzov Lilja Øvrelid Terrence Szymanski[◇] Erik Velldal
University of Oslo, Norway
{andreku | liljao | erikve}@ifi.uio.no
[◇]ANZ, Melbourne, Australia
terry.szymanski@gmail.com

Abstract

Recent years have witnessed a surge of publications aimed at tracing temporal changes in lexical semantics using distributional methods, particularly prediction-based word embedding models. However, this vein of research lacks the cohesion, common terminology and shared practices of more established areas of natural language processing. In this paper, we survey the current state of academic research related to diachronic word embeddings and semantic shifts detection. We start with discussing the notion of semantic shifts, and then continue with an overview of the existing methods for tracing such time-related shifts with word embedding models. We propose several axes along which these methods can be compared, and outline the main challenges before this emerging subfield of NLP, as well as prospects and possible applications.

1 Introduction

The meanings of words continuously change over time, reflecting complicated processes in language and society. Examples include both changes to the core meaning of words (like the word *gay* shifting from meaning 'carefree' to 'homosexual' during the 20th century) and subtle shifts of cultural associations (like *Iraq* or *Syria* being associated with the concept of 'war' after armed conflicts had started in these countries). Studying these types of changes in meaning enables researchers to learn more about human language and to extract temporal-dependent data from texts.

The availability of large corpora and the development of computational semantics have given rise to a number of research initiatives trying to capture *diachronic semantic shifts* in a data-driven way. Recently, *word embeddings* (Mikolov et al., 2013b) have become a widely used input representation for this task. There are dozens of papers on the topic, mostly published after 2011 (we survey them in Section 3 and further below). However, this emerging field is highly heterogenous. There are at least three different research communities interested in it: natural language processing (and computational linguistics), information retrieval (and computer science in general), and political science. This is reflected in the terminology, which is far from being standardized. One can find mentions of 'temporal embeddings,' 'diachronic embeddings,' 'dynamic embeddings,' etc., depending on the background of a particular research group. The present survey paper attempts to describe this diversity, introduce some axes of comparison and outline main challenges which the practitioners face. Figure 1 shows the timeline of events that influenced the research in this area: in the following sections we cover them in detail.

This survey is restricted in scope to research which traces semantic shifts using distributional word embedding models (that is, representing lexical meaning with dense vectors produced from co-occurrence data). We only briefly mention other data-driven approaches also employed to analyze temporal-labeled corpora (for example, topic modeling). Also, we do not cover syntactic shifts and other changes in the functions rather than meaning of words.

The paper is structured as follows. In Section 2 we introduce the notion of 'semantic shift' and provide some linguistic background for it. Section 3 aims to compare different approaches to the task of

Proceedings of the 27th International Conference on Computational Linguistics, pages 1384–1397
Santa Fe, New Mexico, USA, August 20-26, 2018.

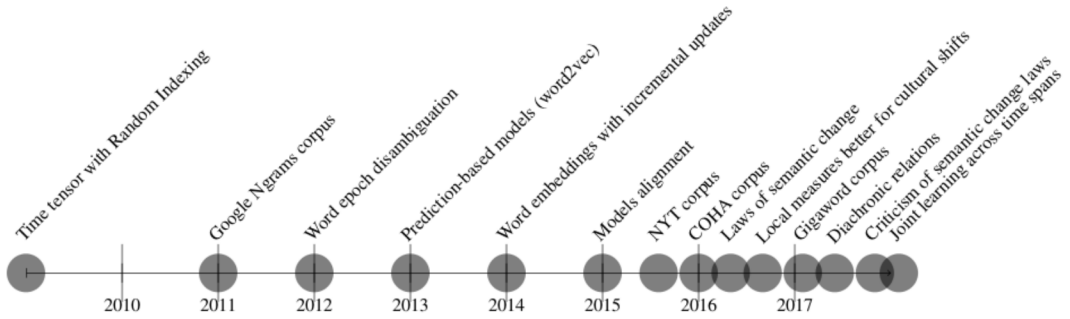

Figure 1: Distributional models in the task of tracing diachronic semantic shifts: research timeline

automatic detection of semantic shifts: in the choice of diachronic data, evaluation strategies, methodology of extracting semantic shifts from data, and the methods to compare word vectors across time spans. Sections 4 and 5 describe two particularly interesting results of diachronic embeddings research: namely, the statistical laws of semantic change and temporal semantic relations. In Section 6 we outline possible applications of systems that trace semantic shifts. Section 7 presents open challenges which we believe to be most important for the field, and in Section 8 we summarize and conclude.

2 The concept of semantic shifts

Human languages change over time, due to a variety of linguistic and non-linguistic factors and at all levels of linguistic analysis. In the field of theoretical (diachronic) linguistics, much attention has been devoted to expressing regularities of linguistic change. For instance, laws of phonological change have been formulated (e.g., Grimm's law or the great vowel shift) to account for changes in the linguistic sound system. When it comes to lexical semantics, linguists have studied the evolution of word meaning over time, describing so-called lexical *semantic shifts* or *semantic change*, which Bloomfield (1933) defines as "innovations which change the lexical meaning rather than the grammatical function of a form."

Historically, much of the theoretical work on semantic shifts has been devoted to documenting and categorizing various types of semantic shifts (Bréal, 1899; Stern, 1931; Bloomfield, 1933). The categorization found in Bloomfield (1933) is arguably the most used and has inspired a number of more recent studies (Blank and Koch, 1999; Geeraerts, 1997; Traugott and Dasher, 2001). Bloomfield (1933) originally proposed nine classes of semantic shifts, six of which are complimentary pairs along a dimension. For instance, the pair 'narrowing' – 'broadening' describes the observation that word meaning often changes to become either more specific or more general, e.g. Old English *mete* 'food' becomes English *meat* 'edible flesh,' or that the more general English word *dog* is derived from Middle English *dogge* which described a dog of a particular breed. Bloomfield (1933) also describes change along the spectrum from positive to negative, describing the speaker's attitude as one of either degeneration or elevation, e.g. from Old English *cniht* 'boy, servant' to the more elevated *knight*.

The driving forces of semantic shifts are varied, but include linguistic, psychological, sociocultural or cultural/encyclopedic causes (Blank and Koch, 1999; Grzega and Schoener, 2007). Linguistic processes that cause semantic shifts generally involve the interaction between words of the vocabulary and their meanings. This may be illustrated by the process of ellipsis, whereby the meaning of one word is transferred to a word with which it frequently co-occurs, or by the need for discrimination of synonyms caused by lexical borrowings from other languages. Semantic shifts may be also be caused by changes in the attitudes of speakers or in the general environment of the speakers. Thus, semantic shifts are naturally separated into two important classes: linguistic drifts (slow and regular changes in core meaning of words) and cultural shifts (culturally determined changes in associations of a given word). Researchers studying semantic shifts from a computational point of view have shown the existence of this division empirically (Hamilton et al., 2016c). In the traditional classification of Stern (1931), the semantic shift category of *substitution* describes a change that has a non-linguistic cause, namely that of technologi-

cal progress. This may be exemplified by the word *car* which shifted its meaning from non-motorized vehicles after the introduction of the automobile.

The availability of large corpora have enabled the development of new methodologies for the study of lexical semantic shifts within general linguistics (Traugott, 2017). A key assumption in much of this work is that changes in a word's collocational patterns reflect changes in word meaning (Hilpert, 2008), thus providing a usage-based account of semantics (Gries, 1999). For instance, Kerremans et al. (2010) study the very recent neologism *detweet*, showing the development of two separate usages/meanings for this word ('to delete from twitter,' vs 'to avoid tweeting') based on large amounts of web-crawled data. The usage-based view of lexical semantics aligns well with the assumptions underlying the distributional semantic approach (Firth, 1957) often employed in NLP . Here, the time spans studied are often considerably shorter (decades, rather than centuries) and we find that these distributional methods seem well suited for monitoring the gradual process of meaning change. Gulordava and Baroni (2011), for instance, showed that distributional models capture cultural shifts, like the word *sleep* acquiring more negative connotations related to sleep disorders, when comparing its 1960s contexts to its 1990s contexts.

To sum up, semantic shifts are often reflected in large corpora through change in the context of the word which is undergoing a shift, as measured by co-occurring words. It is thus natural to try to detect semantic shifts automatically, in a 'data-driven' way. This vein of research is what we cover in the present survey. In the following sections, we overview the methods currently used for the automatic detection of semantic shifts and the recent academic achievements related to this problem.

3 Tracing semantic shifts distributionally

Conceptually, the task of discovery of semantic shifts from data can be formulated as follows. Given corpora $[C_1, C_2, ...C_n]$ containing texts created in time periods $[1, 2, ...n]$, the task is to locate words with different meaning in different time periods, or to locate the words which changed most. Other related tasks are possible: discovering general trends in semantic shifts (see Section 4) or tracing the dynamics of the relationships between words (see Section 5). In the next subsections, we address several axes along which one can categorize the research on detecting semantic shifts with distributional models.

3.1 Sources of diachronic data for training and testing

When automatically detecting semantic shifts, the types of generalizations we will be able to infer are influenced by properties of the textual data being used, such as the source of the datasets and the temporal granularity of the data. In this subsection we discuss the data choices made by researchers (of course, not pretending to cover the whole range of the diachronic corpora used).

3.1.1 Training data

The time unit (the granularity of the temporal dimension) can be chosen before slicing the text collection into subcorpora. Earlier works dealt mainly with long-term semantic shifts (spanning decades or even centuries), as they are easier to trace. One of the early examples is Sagi et al. (2011) who studied differences between Early Middle, Late Middle and Early Modern English, using the Helsinki Corpus (Rissanen and others, 1993).

The release of the Google Books Ngrams corpus[1] played an important role in the development of the field and spurred work on the new discipline of 'culturomics,' studying human culture through digital media (Michel et al., 2011). Mihalcea and Nastase (2012) used this dataset to detect differences in word usage and meaning across 50-years time spans, while Gulordava and Baroni (2011) compared word meanings in the 1960s and in the 1990s, achieving good correlation with human judgments. Unfortunately, Google Ngrams is inherently limited in that it does not contain full texts. However, for many cases, this corpus was enough, and its usage as the source of diachronic data continued in Mitra et al. (2014) (employing syntactic ngrams), who detected word sense changes over several different time periods spanning from 3 to 200 years.

[1] https://books.google.com/ngrams

In more recent work, time spans tend to decrease in size and become more granular. In general, corpora with smaller time spans are useful for analyzing socio-cultural semantic shifts, while corpora with longer spans are necessary for the study of linguistically motivated semantic shifts. As researchers are attempting to trace increasingly subtle cultural semantic shifts (more relevant for practical tasks), the granularity of time spans is decreasing: for example, Kim et al. (2014) and Liao and Cheng (2016) analyzed the *yearly* changes of words. Note that, instead of using granular 'bins', time can also be represented as a continuous differentiable value (Rosenfeld and Erk, 2018).

In addition to the Google Ngrams dataset (with granularity of 5 years), Kulkarni et al. (2015) used Amazon Movie Reviews (with granularity of 1 year) and Twitter data (with granularity of 1 month). Their results indicated that computational methods for the detection of semantic shifts can be robustly applied to time spans less than a decade. Zhang et al. (2015) used another yearly text collection, the New-York Times Annotated Corpus (Sandhaus, 2008), again managing to trace subtle semantic shifts. The same corpus was employed by Szymanski (2017), with 21 separate models, one for each year from 1987 to 2007, and to some extent by Yao et al. (2018), who crawled the NYT web site to get 27 yearly subcorpora (from 1990 to 2016). The inventory of diachronic corpora used in tracing semantic shifts was expanded by Eger and Mehler (2016), who used the Corpus of Historical American (COHA[2]), with time slices equal to one decade. Hamilton et al. (2016a) continued the usage of COHA (along with the Google Ngrams corpus). Kutuzov et al. (2017b) started to employ the yearly slices of the English Gigaword corpus (Parker et al., 2011) in the analysis of cultural semantic drift related to armed conflicts.

3.1.2 Test sets

Diachronic corpora are needed not only as a source of *training* data for developing semantic shift detection systems, but also as a source of *test* sets to evaluate such systems. In this case, however, the situation is more complicated. Ideally, diachronic approaches should be evaluated on human-annotated lists of semantically shifted words (ranked by the degree of the shift). However, such gold standard data is difficult to obtain, even for English, let alone for other languages. General linguistics research on language change like that of Traugott and Dasher (2001) and others usually contain only a small number of hand-picked examples, which is not sufficient to properly evaluate an automatic unsupervised system.

Various ways of overcoming this problem have been proposed. For example, Mihalcea and Nastase (2012) evaluated the ability of a system to detect the time span that specific contexts of a word undergoing a shift belong to (*word epoch disambiguation*). A similar problem was offered as SemEval-2015 Task 7: 'Diachronic Text Evaluation' (Popescu and Strapparava, 2015). Another possible evaluation method is so-called cross-time alignment, where a system has to find equivalents for certain words in different time periods (for example, 'Obama' in 2015 corresponds to 'Trump' in 2017). There exist several datasets containing such temporal equivalents for English (Yao et al., 2018). Yet another evaluation strategy is to use the detected diachronic semantic shifts to trace or predict real-world events like armed conflicts (Kutuzov et al., 2017b). Unfortunately, all these evaluation methods still require the existence of large manually annotated semantic shift datasets. The work to properly create and curate such datasets is in its infancy.

One reported approach to avoid this requirement is borrowed from research on word sense disambiguation and consists of making a synthetic task by merging two real words together and then modifying the training and test data according to a predefined sense-shifting function. Rosenfeld and Erk (2018) successfully employed this approach to evaluate their system; however, it still operates on synthetic words, limiting the ability of this evaluation scheme to measure the models' performance with regards to real semantic shift data. Thus, the problem of evaluating semantic shift detection approaches is far from being solved, and practitioners often rely on self-created test sets, or even simply manually inspecting the results.

3.2 Methodology of extracting semantic shifts from data

After settling on a diachronic data set to be used in the system, one has to choose the methods to analyze it. Before the broad adoption of word embedding models, it was quite common to use change in raw

[2] http://corpus.byu.edu/coha/

word frequencies in order to trace semantic shifts or other kinds of linguistic change; see, among others, Juola (2003), Hilpert and Gries (2009), Michel et al. (2011), Lijffijt et al. (2012), or Choi and Varian (2012) for frequency analysis of words in web search queries. Researchers also studied the increase or decrease in the frequency of a word A collocating with another word B over time, and based on this inferred changes in the meaning of A (Heyer et al., 2009).

However, it is clear that semantic shifts are not always accompanied with changes in word frequency (or this connection may be very subtle and non-direct). Thus, if one were able to more directly model word meaning, such an approach should be superior to frequency-proxied methods. A number of recent publications have showed that *distributional word representations* (Turney et al., 2010; Baroni et al., 2014) provide an efficient way to solve these tasks. They represent meaning with sparse or dense (embedding) vectors, produced from word co-occurrence counts. Although conceptually the source of the data for these models is still word frequencies, they 'compress' this information into continuous lexical representations which are both efficient and convenient to work with. Indeed, Kulkarni et al. (2015) explicitly demonstrated that distributional models outperform the frequency-based methods in detecting semantic shifts. They managed to trace semantic shifts more precisely and with greater explanatory power. One of the examples from their work is the semantic evolution of the word *gay*: through time, its nearest semantic neighbors changed, manifesting the gradual move away from the sense of 'cheerful' to the sense of 'homosexual.'

In fact, distributional models were being used in diachronic research long before the paper of Kulkarni et al. (2015), although there was no rigorous comparison to the frequentist methods. Already in 2009, it was proposed that one can use distributional methods to detect semantic shifts in a quantitative way. The pioneering work by Jurgens and Stevens (2009) described an insightful conceptualization of a sequence of distributional model updates through time: it is effectively a Word:Semantic Vector:Time tensor, in the sense that each word in a distributional model possesses a set of semantic vectors for each time span we are interested in. It paved the way for quantitatively comparing not only words with regard to their meaning, but also different stages in the development of word meaning over time.

Jurgens and Stevens (2009) employed the *Random Indexing* (RI) algorithm (Kanerva et al., 2000) to create word vectors. Two years later, Gulordava and Baroni (2011) used explicit count-based models, consisting of sparse co-occurrence matrices weighted by Local Mutual Information, while Sagi et al. (2011) turned to Latent Semantic Analysis (Deerwester et al., 1990). In Basile et al. (2014), an extension to RI dubbed *Temporal Random Indexing* (TRI) was proposed. However, no quantitative evaluation of this approach was offered (only a few hand-picked examples based on the Italian texts from the *Gutenberg Project*), and thus it is unclear whether TRI is any better than other distributional models for the task of semantic shift detection.

Further on, the diversity of the employed methods started to increase. For example, Mitra et al. (2014) analyzed clusters of the word similarity graph in the subcorpora corresponding to different time periods. Their distributional model consisted of lexical nodes in the graphs connected with weighted edges. The weights corresponded to the number of shared most salient syntactic dependency contexts, where saliency was determined by co-occurrence counts scaled by Mutual Information (MI). Importantly, they were able to detect not only the mere fact of a semantic shift, but also its type: the birth of a new sense, splitting of an old sense into several new ones, or merging of several senses into one. Thus, this work goes into a much less represented class of 'fine-grained' approaches to semantic shift detection. It is also important that Mitra et al. (2014) handle natively the issue of polysemous words, putting the much-neglected problem of word senses in the spotlight.

The work of Kim et al. (2014) was seminal in the sense that it is arguably the first one employing *prediction-based word embedding models* to trace diachronic semantic shifts. Particularly, they used incremental updates (see below) and Continuous Skipgram with negative sampling (SGNS) (Mikolov et al., 2013a).[3] Hamilton et al. (2016a) showed the superiority of SGNS over explicit PPMI-based distributional models in semantic shifts analysis, although they noted that low-rank SVD approximations (Bullinaria and Levy, 2007) can perform on par with SGNS, especially on smaller datasets. Since then,

[3] Continuous Bag-of-Words (CBOW) from the same paper is another popular choice for learning semantic vectors.

the majority of publications in the field started using dense word representations: either in the form of SVD-factorized PPMI matrices, or in the form of prediction-based shallow neural models like SGNS.[4]

There are some works employing other distributional approaches to semantic shifts detection. For instance, there is a strong vein of research based on dynamic topic modeling (Blei and Lafferty, 2006; Wang and McCallum, 2006), which learns the evolution of topics over time. In Wijaya and Yeniterzi (2011), it helped solve a typical digital humanities task of finding traces of real-world events in the texts. Heyer et al. (2016) employed topic analysis to trace the so-called 'context volatility' of words. In the political science, topic models are also sometimes used as proxies to social trends developing over time: for example, Mueller and Rauh (2017) employed LDA to predict timing of civil wars and armed conflicts. Frermann and Lapata (2016) drew on these ideas to trace diachronic word senses development. But most scholars nowadays seem to prefer parametric distributional models, particularly prediction-based embedding algorithms like SGNS, CBOW or GloVe (Pennington et al., 2014). Following their widespread adoption in NLP in general, they have become the dominant representations for the analysis of diachronic semantic shifts as well.

3.3 Comparing vectors across time

It is rather straightforward to train separate word embedding models using time-specific corpora containing texts from several different time periods. As a consequence, these models are also time-specific. However, it is not that straightforward to compare word vectors across different models.

It usually does not make sense to, for example, directly calculate cosine similarities between embeddings of one and the same word in two different models. The reason is that most modern word embedding algorithms are inherently stochastic and the resulting embedding sets are invariant under rotation. Thus, even when trained on the same data, separate learning runs will produce entirely different numerical vectors (though with roughly the same pairwise similarities between vectors for particular words). This is expressed even stronger for models trained on different corpora. It means that even if word meaning is completely stable, the direct cosine similarity between its vectors from different time periods can still be quite low, simply because the random initializations of the two models were different. To alleviate this, Kulkarni et al. (2015) suggested that before calculating similarities, one should first *align* the models to fit them in one vector space, using linear transformations preserving general vector space structure. After that, cosine similarities across models become meaningful and can be used as indicators of semantic shifts. They also proposed constructing the time series of a word embedding over time, which allows for the detection of 'bursts' in its meaning with the *Mean Shift* model (Taylor, 2000). Notably, almost simultaneously the idea of aligning diachronic word embedding models using a distance-preserving projection technique was proposed by Zhang et al. (2015). Later, Zhang et al. (2016) expanded on this by adding the so called 'local anchors': that is, they used both linear projections for the whole models and small sets of nearest neighbors for mapping the query words to their correct temporal counterparts.

Instead of aligning their diachronic models using linear transformations, Eger and Mehler (2016) compared word meaning using so-called 'second-order embeddings,' that is, the vectors of words' similarities to all other words in the shared vocabulary of all models. This approach does not require any transformations: basically, one simply analyzes the word's position compared to other words. At the same time, Hamilton et al. (2016a) and Hamilton et al. (2016c) showed that these two approaches can be used simultaneously: they employed both 'second order embeddings' and orthogonal Procrustes transformations to align diachronic models.

Recently, it was shown in Bamler and Mandt (2017) ('*dynamic skip-gram*' model) and Yao et al. (2018) ('*dynamic Word2Vec*' model) that it is possible to learn the word embeddings across several time periods jointly, enforcing alignment across all of them simultaneously, and positioning all the models in the same vector space in one step. This develops the idea of model alignment even further and eliminates the need to first learn separate embeddings for each time period, and then align subsequent model pairs. Bamler and Mandt (2017) additionally describe two variations of their approach: a) for the cases when data slices arrive sequentially, as in streaming applications, where one can not use future observations, and b) for

[4] Levy and Goldberg (2014) showed that these two approaches are equivalent from the mathematical point of view.

the cases when data slices are available all at once, allowing for training on the whole sequence from the very beginning. A similar approach is taken by Rosenfeld and Erk (2018) who train a deep neural network on word and time representations. Word vectors in this setup turn into linear transformations applied to a continuous time variable, and thus producing an embedding of word w at time t.

Yet another way to make the models comparable is made possible by the fact that prediction-based word embedding approaches (as well as RI) allow for incremental updates of the models with new data without any modifications. This is not the case for the traditional explicit count-based algorithms, which usually require a computationally expensive dimensionality reduction step. Kim et al. (2014) proposed the idea of *incrementally updated diachronic embedding models*: that is, they train a model on the year y_i, and then the model for the year y_{i+1} is initialized with the word vectors from y_i. This can be considered as an alternative to model alignment: instead of aligning models trained from scratch on different time periods, one starts with training a model on the diachronically first period, and then updates this same model with the data from the successive time periods, saving its state each time. Thus, all the models are inherently related to each other, which, again, makes it possible to directly calculate cosine similarities between the same word in different time period models, or at least makes the models more comparable.

Several works have appeared recently which aim to address the technical issues accompanying this approach of incremental updating. Among others, Peng et al. (2017) described a novel method of incrementally learning the *hierarchical softmax* function for the CBOW and Continuous Skipgram algorithms. In this way, one can update word embedding models with new data and new vocabulary much more efficiently, achieving faster training than when doing it from scratch, while at the same time preserving comparable performance. Continuing this line of research, Kaji and Kobayashi (2017) proposed a conceptually similar incremental extension for *negative sampling*, which is a method of training examples selection, widely used with prediction-based models as a faster replacement for *hierarchical softmax*.

Even after the models for different time periods are made comparable in this or that way, one still has to choose the exact method of comparing word vectors across these models. Hamilton et al. (2016a) and Hamilton et al. (2016c) made an important observation that the distinction between linguistic and cultural semantic shifts is correlated with the distinction between *global* and *local* embedding comparison methods. The former take into account the whole model (for example, 'second-order embeddings,' when we compare the word's similarities to all other words in the lexicon), while the latter focus on the word's immediate neighborhood (for example, when comparing the lists of k nearest neighbors). They concluded that global measures are sensitive to regular processes of linguistic shifts, while local measures are better suited to detect slight cultural shifts in word meaning. Thus, the choice of particular embedding comparison approach should depend on what type of semantic shifts one seeks to detect.

4 Laws of semantic change

The use of diachronic word embeddings for studying the dynamics of word meaning has resulted in several hypothesized 'laws' of semantic change. We review some of these law-like generalizations below, before finally describing a study that questions their validity.

Dubossarsky et al. (2015) experimented with K-means clustering applied to SGNS embeddings trained for evenly sized yearly samples for the period 1850–2009. They found that the degree of semantic change for a given word – quantified as the change in self-similarity over time – negatively correlates with its distance to the centroid of its cluster. They proposed that the likelihood for semantic shift correlates with the degree of prototypicality (the *'law of prototypicality'* in Dubossarsky et al. (2017)).

Another relevant study is reported by Eger and Mehler (2016), based on two different graph models; one being a time-series model relating embeddings across time periods to model semantic shifts and the other modeling the self-similarity of words across time. Experiments were performed with time-indexed historical corpora of English, German and Latin, using time-periods corresponding to decades, years and centuries, respectively. To enable comparison of embeddings across time, second-order embeddings encoding similarities to other words were used, as described in 3.3, limited to the 'core vocabulary' (words occurring at least 100 times in all time periods). Based on linear relationships observed in the graphs, Eger and Mehler (2016) postulate two 'laws' of semantic change:

1. word vectors can be expressed as linear combinations of their neighbors in previous time periods;

2. the meaning of words tend to decay linearly in time, in terms of the similarity of a word to itself; this is in line with the '*law of differentiation*' proposed by Xu and Kemp (2015).

In another study, Hamilton et al. (2016a) considered historical corpora for English, German, French and Chinese, spanning 200 years and using time spans of decades. The goal was to investigate the role of frequency and polysemy with respect to semantic shifts. As in Eger and Mehler (2016), the rate of semantic change was quantified by self-similarity across time-points (with words represented by Procrustes-aligned SVD embeddings). Through a regression analysis, Hamilton et al. (2016a) investigated how the change rates correlate with frequency and polysemy, and proposed another two 'laws':

1. frequent words change more slowly ('*the law of conformity*');

2. polysemous words (controlled for frequency) change more quickly ('*the law of innovation*').

Azarbonyad et al. (2017) showed that these laws (at least the law of conformity) hold not only for diachronic corpora, but also for other 'viewpoints': for example, semantic shifts across models trained on texts produced by different political actors or written in different genres (Kutuzov et al., 2016). However, the temporal dimension allows for a view of the corpora under analysis as a sequence, making the notion of 'semantic shift' more meaningful.

Later, Dubossarsky et al. (2017) questioned the validity of some of these proposed 'laws' of semantic change. In a series of replication and control experiments, they demonstrated that some of the regularities observed in previous studies are largely artifacts of the models used and frequency effects. In particular, they considered 10-year bins comprising equally sized yearly samples from Google Books 5-grams of English fiction for the period 1990–1999. For control experiments, they constructed two additional data sets; one with chronologically shuffled data where each bin contains data from all decades evenly distributed, and one synchronous variant containing repeated random samples from the year 1999 alone. Any measured semantic shifts within these two alternative data sets would have to be due to random sampling noise.

Dubossarsky et al. (2017) performed experiments using raw co-occurrence counts, PPMI weighted counts, and SVD transformations (Procrustes aligned), and conclude that the 'laws' proposed in previous studies – that semantic change is correlated with frequency, polysemy (Hamilton et al., 2016a) and prototypicality (Dubossarsky et al., 2015) – are not valid as they are also observed in the control conditions. Dubossarsky et al. (2017) suggested that these spurious effects are instead due to the type of word representation used – count vectors – and that semantic shifts must be explained by a more diverse set of factors than distributional ones alone. Thus, the discussion on the existence of the 'laws of semantic change' manifested by distributional trends is still open.

5 Diachronic semantic relations

Word embedding models are known to successfully capture complex *relationships* between concepts, as manifested in the well-known word analogies task (Mikolov et al., 2013a), where a model must 'solve' equations of the form 'A is to B is as C is to what?' A famous example is the distributional model capturing the fact that the relation between '*man*' and '*woman*' is the same as between '*king*' and '*queen*' (by adding and subtracting the corresponding word vectors). Thus, it is a natural development to investigate whether changes in semantic relationships across time can also be traced by looking at the diachronic development of distributional models.

Zhang et al. (2015) considered the *temporal correspondences problem*, wherein the objective is to identify the word in a target time period which corresponds to a query term in the source time period (for example, given the query term *iPod* in the 2000s, the counterpart term in the 1980s time period is *Walkman*). This is proposed as a means to improve the results of information retrieval from document collections with significant time spans. Szymanski (2017) frames this as the *temporal word analogy* problem, extending the word analogies concept into the temporal dimension. This work shows that

diachronic word embeddings can successfully model relations like 'word w_1 at time period t_α is like word w_2 at time period t_β'. To this end, embedding models trained on different time periods are aligned using linear transformations. Then, the temporal analogies are solved by simply finding out which word vector in the time period t_β is the closest to the vector of w_1 in the time period t_α.

A variation of this task was studied in Rosin et al. (2017), where the authors learn the relatedness of words over time, answering queries like 'in which time period were the words *Obama* and *president* maximally related'. This technique can be used for a more efficient user query expansion in general-purpose search engines. Kutuzov et al. (2017a) modeled a different semantic relation: 'words w_1 and w_2 at time period t_α are in the same semantic relation as words w_3 and w_4 at time period t_β'. To trace the temporal dynamics of these relations, they re-applied linear projections learned on sets of w_1 and w_2 pairs from the model for the period t_n to the model trained on the subsequent time period t_{n+1}. This was used to solve the task of detecting lasting or emerging armed conflicts and the violent groups involved in these conflicts.

6 Applications

Applications of diachronic word embeddings approaches can generally be grouped into two broad categories: *linguistic studies* which investigate the how and why of semantic shifts, and *event detection* approaches which mine text data for actionable purposes.

The first category generally involves corpora with longer time spans, since linguistic changes happen at a relatively slow pace. Some examples falling into this category include tracking semantic drift of particular words (Kulkarni et al., 2015) or of word sentiment (Hamilton et al., 2016b), identifying the breakpoints between epochs (Sagi et al., 2011; Mihalcea and Nastase, 2012), studying the laws of semantic change at scale (Hamilton et al., 2016c) and finding different words with similar meanings at different points in time (Szymanski, 2017). This has been held up as a good use case of deep learning for research in computational linguistics (Manning, 2015), and there are opportunities for future work applying diachronic word embeddings not only in the field of historical linguistics, but also in related areas like sociolinguistics and digital humanities.

The second category involves mining texts for cultural semantic shifts (usually on shorter time spans) indicating real-world events. Examples of this category are temporal information retrieval (Rosin et al., 2017), predicting civil turmoils (Kutuzov et al., 2017b; Mueller and Rauh, 2017), or tracing the popularity of entities using norms of word vectors (Yao et al., 2018). They can potentially be employed to improve user experience in production systems or for policy-making in governmental structures.

We believe that the near future will see a more diverse landscape of applications for diachronic word embeddings, especially related to the real-time analysis of large-scale news streams. 'Between the lines,' these data sources contain a tremendous amount of information about processes in our world, manifested in semantic shifts of various sorts. The task of researchers is to reveal this information and make it reliable and practically useful.

7 Open challenges

The study of temporal aspects of semantic shifts using distributional models (including word embeddings) is far from being a solved problem. The field still has a considerable number of open challenges. Below we briefly describe the most demanding ones.

- The existing methods should be expanded to a *wider scope of languages*. Hamilton et al. (2016a), Kutuzov and Kuzmenko (2018) and others have started to analyze other languages, but the overwhelming majority of publications still apply only to English corpora. It might be the case that the best methodologies are the same for different languages, but this should be shown empirically.

- There is a clear need to devise algorithms that work on *small datasets*, as they are very common in historical linguistics, digital humanities, and similar disciplines.

- Carefully designed and robust *gold standard test sets* of semantic shifts (of different kinds) should be created. This is a difficult task in itself, but the experience from synchronic word embeddings evaluation (Hill et al., 2015) and other NLP areas proves that it is possible.

- There is a need for rigorous *formal mathematical models of diachronic embeddings*. Arguably, this will follow the vein of research in joint learning across several time spans, started by Bamler and Mandt (2017) and Yao et al. (2018), but other directions are also open.

- Most current studies stop after stating the simple fact that a semantic shift has occurred. However, more detailed analysis of the nature of the shift is needed. This includes:

    1. *Sub-classification of types of semantic shifts* (broadening, narrowing, etc). This problem was to some degree addressed by Mitra et al. (2014), but much more work is certainly required to empirically test classification schemes proposed in much of the theoretical work described in Section 2.
    2. *Identifying the source of a shift* (for example, linguistic or extra-linguistic causes). This causation detection is closely linked to the division between linguistic drifts and cultural shifts, as proposed in Hamilton et al. (2016c).
    3. *Quantifying the weight of senses* acquired over time. Many words are polysemous, and the relative importance of senses is flexible (Frermann and Lapata, 2016). The issue of handling senses is central for detecting semantic shifts, but most of the algorithms described in this survey are not sense-aware. To address this, methods from sense embeddings research (Bartunov et al., 2016) might be employed.
    4. *Identifying groups of words that shift together* in correlated ways. Some work in this direction was started in Dubossarsky et al. (2016), who showed that verbs change more than nouns, and nouns change more than adjectives. This is also naturally related to proving the (non-)existence of the 'laws of semantic change' (see Section 4).

- Last but not least, we believe that the community around diachronic word embeddings research severely lacks relevant forums, like *topical workshops* or *shared tasks*. Diachronic text evaluation tasks like the one at *SemEval-2015* (Popescu and Strapparava, 2015) are important but not enough, since they focus on identifying the time period when a text was authored, not the process of shifting meanings of a word. Organizing such events can promote the field and help address many of the challenges described above.

8 Summary

We have presented an outline of the current research related to computational detection of semantic shifts using diachronic (temporal) word embeddings. We covered the linguistic nature of semantic shifts, the typical sources of diachronic data and the distributional approaches used to model it, from frequentist methods to contemporary prediction-based models. To sum up, Figure 1 shows the timeline of events that have been influential in the development of research in this area: introducing concepts, usage of corpora and important findings.

This emerging field is still relatively new, and although recent years has seen a string of significant discoveries and academic interchange, much of the research still appears slightly fragmented, not least due to the lack of dedicated venues like workshops, special issues, or shared tasks. We hope that this survey will be useful to those who want to understand how this field has developed, and gain an overview of what defines the current state-of-the-art and what challenges lie ahead.

Acknowledgements

We thank William Hamilton, Haim Dubossarsky and Chris Biemann for their helpful feedback during the preparation of this survey. All possible mistakes remain the sole responsibility of the authors.

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the ACM on Conference on Information and Knowledge Management*, pages 1509–1518, Singapore.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the International Conference on Machine Learning*, pages 380–389, Sydney, Australia.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, Baltimore, USA.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 130–138, Cadiz, Spain.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *Proceedings of the First Italian Conference on Computational Linguistics*, pages 38–42, Turin, Italy.

Andreas Blank and Peter Koch. 1999. *Historical semantics and cognition*. Walter de Gruyter.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning*, pages 113–120, Pittsburgh, USA.

Leonard Bloomfield. 1933. *Language*. Allen & Unwin.

Michel Bréal. 1899. *Essai de sémantique*. Hachette, Paris.

John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google trends. *Economic Record*, 88(s1):2–9.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of the NetWordS 2015 Word Knowledge and Word Usage*, pages 66–70, Pisa, Italy.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2016. Verbs change more than nouns: a bottom-up computational approach to semantic change. *Lingue e linguaggio*, 15(1):7–28.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 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*, pages 1147–1156, Copenhagen, Denmark.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 52–58, Berlin, Germany.

John Firth. 1957. *A synopsis of linguistic theory, 1930-1955*. Blackwell.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association of Computational Linguistics*, 4:31–45.

Dirk Geeraerts. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Clarendon Press, Oxford.

Stefan Th. Gries. 1999. Particle movement: a cognitive and functional approach. *Cognitive Linguistics*, 10:105–145.

Joachim Grzega and Marion Schoener. 2007. English and general historical lexicology. *Eichstätt-Ingolstadt: Katholische Universität*.

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, Edinburgh, UK.

L. William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016b. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016c. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas.

Gerhard Heyer, Florian Holz, and Sven Teresniak. 2009. Change of topics over time – tracking topics by their change of meaning. In *Proceeding of the International Conference on Knowledge Discovery and Information Retrieval*, pages 223–228, Madeira, Portugal.

Gerhard Heyer, Cathleen Kantner, Andreas Niekler, Max Overbeck, and Gregor Wiedemann. 2016. Modeling the dynamics of domain specific terminology in diachronic corpora. In *Proceedings of the 12th International conference on Terminology and Knowledge Engineering (TKE 2016)*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Martin Hilpert and Stefan Th. Gries. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

M. Hilpert. 2008. *Germanic future constructions: A usage-based approach to language change*. Benjamins, Amsterdam, Netherlands.

Patrick Juola. 2003. The time course of language change. *Computers and the Humanities*, 37(1):77–96.

David Jurgens and Keith Stevens. 2009. Event detection in blogs using Temporal Random Indexing.

Nobuhiro Kaji and Hayato Kobayashi. 2017. Incremental skip-gram model with negative sampling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Copenhagen, Denmark.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, volume 1036, pages 103–106, Mahwah, USA.

D. Kerremans, S. Stegmayr, and H.-J. Schmid. 2010. The neocrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In K. Allan and J. A. Robinson, editors, *Current methods in historical semantics*, pages 130–160. De Gruyter Mouton.

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 52nd Annual Meeting of the Association for Computational Linguistics*, pages 61–65, Baltimore, USA.

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*, pages 625–635, Florence, Italy.

Andrey Kutuzov and Elizaveta Kuzmenko. 2018. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. *Quantitative Approaches to the Russian Language*, pages 95–112.

Andrey Kutuzov, Elizaveta Kuzmenko, and Anna Marakasova. 2016. Exploration of register-dependent lexical semantics using word embeddings. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 26–34, Osaka, Japan.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017a. Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1824–1829, Copenhagen, Denmark.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017b. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop at ACL 2017*, pages 31–36, Vancouver, Canada.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2177–2185, Montreal, Canada.

Xuanyi Liao and Guang Cheng. 2016. Analysing the semantic change based on word embedding. In *Natural Language Understanding and Intelligent Applications*, pages 213–223. Springer International Publishing.

Jefrey Lijffijt, Tanja Säily, and Terttu Nevalainen. 2012. CEECing the baseline: Lexical stability and significant change in a historical corpus. In *Studies in Variation, Contacts and Change in English*, volume 10. Research Unit for Variation, Contacts and Change in English (VARIENG).

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

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, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 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*, pages 259–263, Jeju Island, Korea.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1020–1029, Baltimore, Maryland.

Hannes Mueller and Christofer Rauh. 2017. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, page 1–18.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition LDC2011T07. Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

Hao Peng, Jianxin Li, Yangqiu Song, and Yaopeng Liu. 2017. Incrementally learning the hierarchical softmax function for neural language models. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3267–327, San Francisco, California USA.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Octavian Popescu and Carlo Strapparava. 2015. SemEval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878, Denver, Colorado.

Matti Rissanen et al. 1993. The helsinki corpus of english texts. *Kyttö et. al*, pages 73–81.

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*, pages 474–484, New Orleans, Louisiana, USA.

Guy D. Rosin, Eytan Adar, and Kira Radinsky. 2017. Learning word relatedness over time. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1179–1189, Copenhagen, Denmark.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, pages 161–183.

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

Gustaf Stern. 1931. *Meaning and change of meaning; with special reference to the English language.* Wettergren & Kerbers.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 448–453, Vancouver, Canada.

Wayne A Taylor. 2000. Change-point analysis: a powerful new tool for detecting changes.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in semantic change*. Cambridge University Press.

Elizabeth Traugott. 2017. Semantic change. *Oxford Research Encyclopedias: Linguistics*.

Peter Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, Philadelphia, PA, USA.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on Detecting and Exploiting Cultural diversity on the social web*, pages 35–40.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Austin, TX, USA.

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*, pages 673–681, Marina Del Rey, CA, USA.

Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 645–655, Beijing, China.

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, October.