Diachronic Analysis of Entities by Exploiting Wikipedia Page revisions

Pierpaolo Basile
University of Bari Aldo Moro
Dept. of Computer Science
Bari, Italy
pierpaolo.basile@uniba.it

Annalina Caputo
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
annalina.caputo@adaptcentre.ie

Seamus Lawless
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
seamus.lawless@adaptcentre.ie

Giovanni Semeraro
University of Bari Aldo Moro
Dept. of Computer Science
Bari, Italy
giovanni.semeraro@uniba.it

Abstract

In the last few years, the increasing availability of large corpora spanning several time periods has opened new opportunities for the diachronic analysis of language. This type of analysis can bring to the light not only linguistic phenomena related to the shift of word meanings over time, but it can also be used to study the impact that societal and cultural trends have on this language change. This paper introduces a new resource for performing the diachronic analysis of named entities built upon Wikipedia page revisions. This resource enables the analysis over time of changes in the relations between entities (concepts), surface forms (words), and the contexts surrounding entities and surface forms, by analysing the whole history of Wikipedia internal links. We provide some useful use cases that prove the impact of this resource on diachronic studies and delineate some possible future usage.

1 Introduction

The availability of large corpora spanning different time periods has encouraged researchers to analyse language from a diachronic perspective. Language is dynamic and detecting significant linguistic shifts in the meaning and usage of words is a crucial task for both social and cultural studies and for Natural Language Processing applications. Recent work focusing on the automatic detection of the semantic shift of words has adopted diachronic (or dynamic) word embeddings (Kim et al., 2014; Hamilton et al., 2016b; Kulkarni et al., 2015). This type of work represents words as vectors in a *semantic* space where the proximity between word vectors indicate the existence of a semantic relationship between the terms involved. The diachronic analysis is then performed by building a different semantic space for each period of investigation and aligning vectors belonging to different spaces in order to make them comparable. The variations in the similarity between the word vectors in two different spaces marks possible changes in the context of appearance of that word. This is used as a proxy indicator of change, either cultural, social or semantic, associated with the occurrence of that specific word. This kind of work has generated a variety of resources for the diachronic analysis of word meanings, covering different time periods, languages, and genres.

While the broader area of automatic detection of semantic shift of words is gaining momentum, only little effort has focused on the more specific problem of analysing the semantic shift of named entities. This problem has a huge impact on the correct identification of entities in context, with repercussions on many natural language processing problems, such as entity linking and search, aspect-based sentiment analysis and event detection (Kanhabua and Nørvåg, 2010b; Tahmasebi et al., 2012; Georgescu et al., 2013).

Generally, an entity has a clear referent and what evolves is the context in which it appears or the surface form used to refer to it. In this work, we build a resource that tracks how the surface forms used to link an entity change over time by taking into account the revisions of Wikipedia pages. In doing so, we also extract time-dependent contexts of each mention of a link in Wikipedia

Proceedings of Recent Advances in Natural Language Processing, pages 84–91,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_011

pages. The Wikipedia page history, sometimes called revision history or edit history, tracks the order in which changes were made to any editable Wikipedia page. We believe that this corpus can help researchers to design approaches for tracking entities usage over time. This resource can be functional to promote new research for dynamic embeddings of named entities. We propose some preliminary case studies for proving the potentiality of this resource.

The paper is structured as follows: Section 2 reviews the state of the art, while Section 3 describes the methodological aspects of our approach. Section 4 shows some use cases of our resource followed by some final remarks.

2 Related Work

The diachronic analysis of language via word embeddings has been an active area of research during the past decade that has generated many resources for several time periods, languages and genres. Kim et al. (2014) used Google Ngram as a diachronic resource to build word embeddings via Word2Vec on a random sample of the 10 million 5-grams from the English fiction portion of the corpus. The authors made the resource available, but due to space limitations, they released the word embeddings only for the 5-year time period. A similar approach was proposed by Grayson et al. (2016), where Word2Vec embeddings are trained on the Eighteenth-Century Collections Online corpus (ECCO-TCP) by taking into account five twenty-year periods for 150 million words randomly sampled from the "Literature and Language" section of the corpus. Hamilton et al. (2016b) also trained word embeddings on the Google Ngram for detecting semantic changes. The authors analysed four different languages, i.e. English, French, German and Chinese, and created a resource which has been successfully used in subsequent studies (Garg et al., 2017; Hamilton et al., 2016a). A different approach to detect the semantic shift of words was adopted by Kulkarni et al. (2015). The authors adopt a change point detection algorithm on the time series generated by computing the cosine similarity between word embeddings trained on several corpora, such as: Twitter, Amazon reviews, and the Google Book Ngrams. A similar approach is proposed in Basile and McGillivray (2018), in which the Temporal Random Indexing (TRI) is adopted for building a distributed, time-based, word representation for the JISC UK Web Domain Dataset (1996-2013) corpus.

Other research efforts have been directed to release resources and applications for the visual analysis and querying of these diachronic collections. The Google Ngram viewer (Michel et al., 2011) was released as a tool for allowing users to query the Google Ngram corpus, a collection of ngram occurrences spanning several years and languages extracted from the Google Book project. Hellrich and Hahn (2017) proposed a system that allows users to explore different corpora via a diachronic semantic search. They used the Corpus of Historical American English, the Deutsches Textarchiv "German Text Archive", and the Royal Society Corpus, in addition to the Google Books Ngram Corpus.

Research directed toward the specific problem of detecting changes in the context surrounding named entities has attracted limited attention compared to the broader area of automatic detection of the semantic shift of words. Some previous work on named entities focused on problems related to searching (Berberich et al., 2009; Kanhabua and Nørvåg, 2010a; Zhang et al., 2016). Tahmasebi et al. (2012) proposed an interesting approach to identify the evolution of named entities. Berberich et al. (2009) defined a method for query reformulations able to paraphrase the user's information need using terminology prevalent in the past. In this work, the original dataset is enriched with annotated phrases extracted from the text by using Wikipedia page titles. In Kanhabua and Nørvåg (2010a), Wikipedia internal links and redirect pages are exploited for finding synonyms across time by using different snapshot of Wikipedia. The identified synonyms are used for query expansion in order to increase the retrieval effectiveness. In some respects, this approach is similar to ours. However, it does not use page revisions and the relation between concepts, surface forms and contexts. Zhang et al. (2016) described an approach to find *past* similar terms closest to a given *present* term. The goal was to improve the retrieval effectiveness in archives and collections of past documents. In this work, Wikipedia is only functional to the creation of the test set, where only the information about the entity lifetime is used (e.g. the time when the name of a country or a company changed). Regarding named entity evolution, Tahmasebi et al. (2012) proposed a method to capture the evolution of one name into another by using a sliding window of co-occurrence terms. The corpus used for the evaluation is the New York Times Annotated Corpus. Lansdall-Welfare et al. (2017) analysed a collection of historical data spanning 150 years of British articles. The authors focus on historical and cultural changes that are tracked via a quantitative analysis of word frequencies. However, they expand their methodology to a "semantic" level by working on named entities extracted from text. The work proposed in Szymanski (2017) is the first attempt to highlight the potential of diachronic word embeddings for solving analogy tasks involving entities and relationships, although this work does not seek to capture named entities in an explicit way. Moreover, Caputo et al. (2018) applied a method to recognise and linking named entities in the whole New York Times corpus. The Temporal Random Indexing is then applied on the annotated corpus in order to build a semantic vector representation for entities and tracking significant shift in their contexts. An explicit representation of named entities is also provided in (Bianchi et al., 2018) where the authors tackle the problem of incorporating time in the Knowledge Graph embeddings in order to provide a similarity measure between entities that accounts for temporal factors.

3 Methodology

The revision history associated with each Wikipedia page opens the way to different diachronic analyses of the highly interconnected concepts represented by its pages. In Wikipedia, pages are interconnected by internal links manually created by users that consist of a surface form and a target. The target is another Wikipedia page, and can be regarded as a "conceptual" link created by the user between the surface form and a specific concept (the Wikipedia page). The same surface form can link several entities and the same entity can be linked to several surface forms. Moreover, since a surface form occurs in a specific context, we can define the surface context as a window of n words to the left and to the right of the surface form. Each page has multiple revisions created every time a user edits that page, and each revision page is associated with a timestamp, so that it is possible to track the changes over time of the temporal relation existing between the surface form, the target and context. For example, it is possible to track the change over time of different surface forms linking to a specific target or to detect the change in the target context. All these capabilities open several possibilities to the analysis of entities using a diachronic perspective.

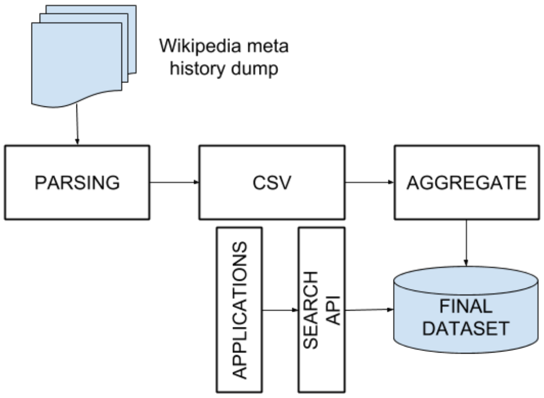


Figure 1: Flowchart of the dataset creation.

Figure 1 depicts the process followed for the creation of our resource. The starting point is the Wikipedia meta history dump which includes all the page revisions in XML format. The dump is composed of several XML files containing the page revisions in Mediawiki syntax. Each XML file is parsed using the DKPro-JWPL API[1], which is able to produce the accurate Abstract Syntax Tree (AST) of each page revision. From the AST, we extract all the internal links that refer to standard[2] Wikipedia pages; each internal link has a surface form and the name of the linked Wikipedia page. In addition, we extract the year from the revision timestamp and the context as the n words around the internal link. The context is processed using the *StandardAnalyzer* provided by the Apache Lucene API[3]. Each extracted internal link is saved in a CSV file as a record consisting of: year, pageId, target, surface form, left context and right context.

An example of a row in a CSV file is reported below:

[1] https://github.com/dkpro/dkpro-jwpl

[2] We remove links to special pages, such as category and user pages.

[3] http://lucene.apache.org/

```
  2003 11057 forge forge forging
term shaping metal use heat
hammer basic smithy contains
sometimes called hearth heating
metals commonly iron steel
malleable temperature
```

The row meaning is that in page *11057* in the year *2003* the target *forge* is linked by the surface form *forge* with the right context *forging, term, ...* and the left context *sometimes, called, ...*.

Since the tuple <*year, surface form, target*> can occur multiple times, we aggregate multiple tuple occurrences in a single record. The aggregation step is performed several times, one time for each dump file plus a final step that aggregates all the records in a single file that represents our final dataset.

In the final file, information is stored as follows:

- A row starting with the sequence *#T* <*TAB*>T_i which identifies the beginning of a sequence of rows in the file that are related to the page (concept) T_i (until a new row starting with *#T* is encountered). T_i represents the Wikipedia page title;

- A sequence of rows containing several values separated by the *tabular* character in the form: year y_k, surface form s_j, the number of time that the surface s_j is used for linking T_i in the year y_k. Then, we build a Bag-of-Word (BoW) from the words occurring in the context, and in the same row we provide the BoW size followed by all the words in the BoW represented as a sequence of pairs <*word, occurrences*>.

A row in the aggregate format is shown in the following example:

```
  #T Apple Computer
2018 Apple Computer 2 30 freedos
1 x 1 supports 1 support 3
officially 1 10 1 s 1 programming
1 9 1 scsi 1 bda 1 2005 1 usb
2 mac 3 announced 2 storage 2
august 1 31 1 ray 2 advanced 1 os
3 its 2 interface 2 blu 2 joined
1 aspi 1 march 1 8.5.1 1 disc 2
mass 2
2018 Apple 1 21 developed 1
computer 1 years 1 independently
1 group 1 computer's 1 1987
1 he 1 while 1 advanced 1
henson 1 associates 1 eric 1
tracking 1 facial 1 technology
1 collaborated 1 six 1 starting 1
worked 1 animation 1
```

The aggregated format shows that the surface form *Apple Computer* was used twice for linking the target *Apple Computer*, while the surface form *Apple* was used only once. The BoW follows each surface form. In the first aggregation step, an aggregated file is created for each segment of the Wikipedia dump, then in the second aggregation step, all the segments are merged in the final dataset.

In this first version of the dataset we do not take into account disambiguation pages and redirects. Managing redirects is a very challenging problem since they are not consistent over dumps.

Relying on this final dataset, we built a search API for easily retrieving all the information related to the target, the surface form and the context according to a specific time period[4].

We exploit the meta history dump dated 1st February 2019; the first Wikipedia pages are dated 2001. The original dump size is about 950GB, the total size of the CSV files is about 30GB, while the final dataset obtained by aggregating data from the CSVs is about 47GB. We set to 10 the dimension of the context window. Since a page can have multiple revisions in the same year, in building our resource we consider only the latest one for each year. It is possible to perform a more fine-grained analysis by taking into account more revisions per year (e.g. a revision for each month). The total number of distinct targets is about 31M, which is larger than the effective number of Wikipedia pages for several reasons: 1) some targets are a redirect to other targets; 2) some pages have been removed or renamed over the years; 3) some targets are a link to a specific section of a page. In this release, we do not take into account these issues, which we plan to tackle in a future release.

The search API can be used for building several applications, such as a RESTful Web Services for remotely querying the data, data analysis for discovering when named entities or surface forms change their usage, and data visualisation.

It is important to underline that the proposed approach is completely unsupervised and language independent since it does not require any NLP pre-

[4] The dataset and the source code are available here https://github.com/pippokill/dae

processing step. Moreover, the proposed methodology is intrinsically multi-language because it is possible to rely on the specific Wikipedia dump of the language under analysis. In addition, it is possible to exploit multi-language Wikipedia links for comparing the evolution of named entities across different languages.

One limit of our approach is the short time frame taken into account since Wikipedia was launched in 2001. However, our approach is incremental and the dataset can grow when new Wikipedia dumps are available. Moreover, the dataset is not only useful for diachronic analysis of entities, but the detection of semantic changes over a short period of time can be exploited to improve the performance of several algorithms, such as entity linking, relation extraction and ontology/knowledge graph population.

4 Use cases

In this section we report some use cases that have emerged from an exploratory analysis of the proposed dataset. We perform the analysis by indexing the 1M most frequent targets extracted from the final dataset. We build an API for querying the dataset by using the Apache Lucene library. Each following subsection reports details about a specific use case.

4.1 Concepts linked by a surface form

The first use case concerns the analysis of the concepts linked by a surface form over time. Table 1 shows the concepts linked by the surface form "Donald Trump". While before 2015 there is only one concept linked by this surface form, since 2016, the concepts related to the presidential campaign have emerged, with the concept "Presidency of Donald Trump" occurring in the first top-5 concepts since 2018. It is important to underline that the first column (2015) reports only one concept since no other concepts are related to the surface form "Donald Trump" in the 2015. This is due to the fact that in this preliminary study we limited our analysis to the 1M most frequent targets and not the whole set of 33M targets. A reverse analysis shows the usage of the surface form "President Trump" to refer to the concept "Donald Trump" since 2017.

4.2 Contexts of a given target

Another interesting analysis concerns the change over time of the contexts of a given target. In this case, it is possible to compute the displacement over time of the target concept by computing the cosine similarity between the context BoWs. For each pair of years, we build a BoW vector for the context of the target concept. Then, we generate a time series by computing the cosine similarity between the BoW of two consecutive years (BoW_{y_i} and $BoW_{y_{i+1}}$). Figure 2 reports the time series generated for the concept "Donald Trump"; we observe a change point corresponding to a drop in similarity between 2015 and 2016.

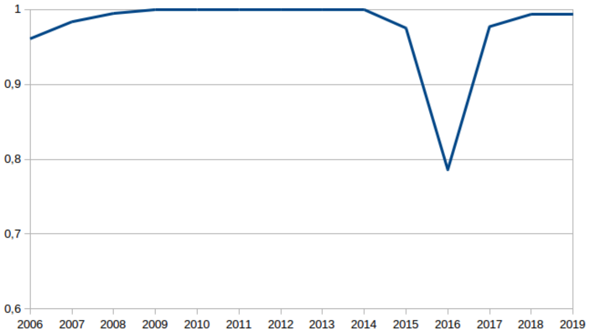


Figure 2: BoW cosine similarity time series for the concept "Donald Trump".

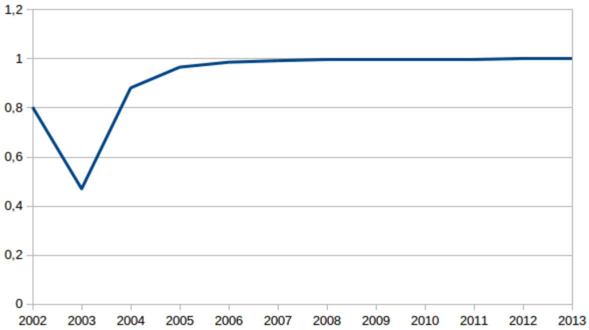


Figure 3: BoW cosine similarity time series for the concept "Arnold Schwarzenegger".

A similar analysis performed for the concept "Arnold Schwarzenegger" shows a change point between 2002-2003 and 2003-2004, as reported in Figure 3. Through the analysis of the most frequent words in the BoWs of the contexts of "Arnold Schwarzenegger" in the period 2002-2004, it emerged that while the most frequent words in 2002 were *film, actor, movie, terminator*, in 2003 new words such as *governor* and *California* related to "Arnold Schwarzenegger" political

2015	2016	2017	2018	2019
Donald Trump	Donald Trump presidential campaign, 2016	Donald Trump	Donald Trump	Donald Trump
	Protests against Donald Trump	Protests against Donald Trump	Protests against Donald Trump	Donald Trump presidential campaign, 2016
	Donald Trump sexual misconduct allegations	Donald Trump presidential campaign, 2016	Donald Trump presidential campaign, 2016	Protests against Donald Trump
	Political positions of Donald Trump	Donald Trump sexual misconduct allegations	Donald Trump sexual misconduct allegations	Donald Trump sexual misconduct allegations
	Stop Trump movement	Donald Trump (Last Week Tonight)	Presidency of Donald Trump	Presidency of Donald Trump

Table 1: Top-5 concepts linked by the surface form “Donald Trump”.

activity have started to appear, to become the most frequent words in the BoWs since 2004.

Another interesting use case is the analysis of the BoWs of the targets linked by the same surface form. This analysis may highlight changes in the way common words are used for referring to named entities. For example, analysing the usage of the surface form “tweet”, we observe that since 2012 it has been used to refer to the concept “Twitter”, while before 2012 it did not refer to any concept.

4.3 Similarity between two entities over time

The last scenario shows the possibility to compute the similarity between two entities over time as the cosine similarity between the target contexts. Figure 4 reports the time series of similarities between three pairs of entities (Apple-Microsoft, Apple-IBM, IBM-Microsoft). It is interesting to observe that the similarity between IBM and Microsoft is higher then the similarity between Apple and the other two entities, although Apple is equally related to both of them.

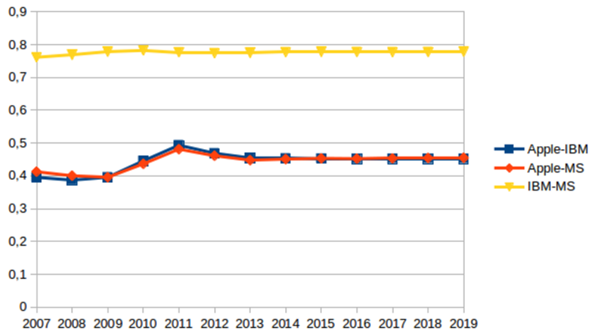


Figure 4: Comparison between pair of entities.

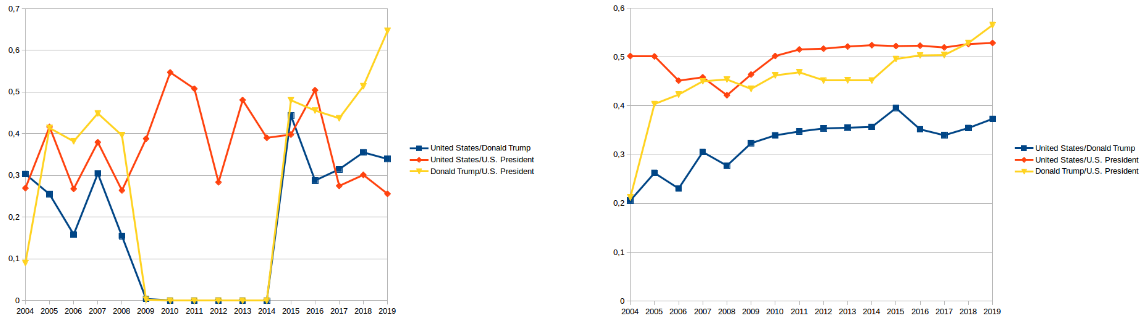
Finally, the plots in Figure 5 show the cosine similarity between the BoWs of two different targets (concepts). Using this approach it is possible to show how the similarity between two targets changes over time. In particular, for each time point we build the BoW of each concept and then we compute the similarity between the BoWs. It is important to point out that the target BoW is built by taking into account the context around each occurrence of the target in the corpus. In this way, if two targets occur in similar contexts their BoWs will be similar. We adopt two strategies:

point-wise: each BoW is built by taking into account only the target occurrences at time t_i;

cumulative: each BoW is built by taking into account all the target occurrences up to time t_i, including time t_i. The idea is to take into account all the previous history of the target and not only the time period under analysis.

Observing the plots in Figure 5, it is possible to note that the similarity between *United States-U.S. President* and *United States-Donald Trump* is constant across time, while we observe an increment in similarity between *U.S. President-Donald Trump* after the year 2018. This increment is clearly evident in the point-wise analysis (Figure 5a), as we expected. It is important to underline that in Figure 5a some points are near zero (2009-2014) this means that the targets do not occur in similar contexts in that periods and indeed the two BoWs share just a few words. Figure 5b show a different trend, since we take into account all the previous target occurrences before the time t_i by exploiting the cumulative approach.

The promising results obtained in this preliminary case study about BoW similarity suggest that it is possible to build effective “time-dependent” embeddings by using our resource.



(a) Plot of the point-wise targets BoW cosine similarity over time.

(b) Plot of the cumulative targets BoW cosine similarity over time.

Figure 5: BoW analysis of pair of targets: plot over time of the cosine similarity between BoWs of two targets with point-wise (a) and cumulative (b) strategy.

5 Conclusions and Future Work

In this paper, we described the construction and utilisation of a new resource built upon Wikipedia page revisions that enables the diachronic analysis of entities. Using the timestamp provided by each revision, we tracked Wikipedia internal links in order to extract the temporal relations between surface forms, contexts, and concepts (Wikipedia pages). We provided some preliminary use cases which show the effectiveness of this resource. These preliminary results show the potentiality of our methodology and open several research scenarios that can be investigated as future work, such as semantic change point detection of entities, entity linking in diachronic collections, event detection, and temporal entity search. The preliminary version of our dataset has some issues that we plan to fix in future versions such as redirects, disambiguation pages and character encoding issues.

Acknowledgement

This work is partially funded by the “TALIA - Territorial Appropriation of Leading-edge Innovation Action” project, Interreg-Mediterranean program for increasing transnational activity of innovative clusters and networks of key sectors of the MED area (2018-2019) and by the ADAPT Centre for Digital Content Technology, funded under the SFI - Science Foundation Ireland Research Centres Programme (Grant SFI 13/RC/2106) and is co-funded under the European Regional Development Fund and by the EU2020 - European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No.: EU2020-713567.

References

Pierpaolo Basile and Barbara McGillivray. 2018. *Discovery Science*, Springer-Verlag, volume 11198 of *Lecture Notes in Computer Science*, chapter Exploiting the Web for Semantic Change Detection.

Klaus Berberich, Srikanta J Bedathur, Mauro Sozio, and Gerhard Weikum. 2009. Bridging the terminology gap in web archive search. In *WebDB*.

Federico Bianchi, Matteo Palmonari, and Debora Nozza. 2018. Towards encoding time in text-based entity embeddings. In Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl, editors, *The Semantic Web – ISWC 2018*. Springer International Publishing, pages 56–71.

Annalina Caputo, Gary Munnelly, and Séamus Lawless. 2018. Temporal entity random indexing. In Jonathan Girón Palau and Isabel Galina Russell, editors, *Digital Humanities 2018, DH 2018, Book of Abstracts, El Colegio de México, UNAM, and RedHD, Mexico City, Mexico, June 26-29, 2018*. Red de Humanidades Digitales A. C., pages 460–461. https://dh2018.adho.org/en/temporal-entity-random-indexing/.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2017. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *PNAS* 115. http://arxiv.org/abs/1711.08412.

Mihai Georgescu, Nattiya Kanhabua, Daniel Krause, Wolfgang Nejdl, and Stefan Siersdorfer. 2013. Extracting event-related information from article updates in wikipedia. In Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 254–266.

Siobhán Grayson, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2016. Novel2Vec

: Characterising 19th Century Fiction via Word Embeddings. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2116–2121. https://doi.org/10.18653/v1/D16-1229.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096* .

Johannes Hellrich and Udo Hahn. 2017. Exploring Diachronic Lexical Semantics with J E S EM E. In *Proceedings of ACL 2017, System Demonstrations*. pages 31–36. http://aclweb.org/anthology/P17-4006.

Nattiya Kanhabua and Kjetil Nørvåg. 2010a. Exploiting time-based synonyms in searching document archives. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, pages 79–88.

Nattiya Kanhabua and Kjetil Nørvåg. 2010b. Quest: Query expansion using synonyms over time. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 595–598.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. *Arxiv* pages 61–65. http://arxiv.org/abs/1405.3515.

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*. ACM, pages 625–635.

Thomas Lansdall-Welfare, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences* 114(4):E457–E465.

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

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 (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 448–453. https://doi.org/10.18653/v1/P17-2071.

Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. Neer: An unsupervised method for named entity evolution recognition. *Proceedings of COLING 2012* pages 2553–2568.

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.