Study concept drift in 150-year English literature

Ruiyuan Li, Pin Tian, and Shenghui Wang

University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands
{r.li-3,p.tian}@student.utwente.nl, shenghui.wang@utwente.nl

Abstract. The meaning of a concept or a word changes over time. Such concept drift reflects the change of the social consensus as well. Studying concept drift over time is valuable for researchers who are interested in language or culture evolution. Recent word embedding technologies inspire us to automatically detect concept drift in large-scale corpora. However, comparing embeddings generated from different corpora is a complex task. In this paper, we propose to use a simple approach for detecting concept drift based on the change in word contexts from different time periods and apply it to subsequent time periods so that the detailed drift could be detected and visualised. We dive into certain words to track how the meaning of a word changes gradually over a long time span with relevant historical events which demonstrates the effect of our method.

Keywords: concept drifting · word embedding · historical event · word context

1 Introduction

Concept drift or diachronic semantic shift studies how the meaning of a concept or a word changes over time [15, 12]. Concept drift reflects the change of the social consensus. For example, the word gay was originally used to mean carefree, cheerful, or bright [7]. However, from the 1960s, the word gay started to describe homosexual men [3]. Studying concept drift over time is valuable for researchers who are interested in language or culture evolution. For people who want to identify societal changes in literature, who research on historical texts, such as librarians, historians or linguists, it is desirable if they can discover potential concept drift in large-scale textual content before conducting in-depth investigation. Automatically identifying concept drift over time can improve their efficiency.

Recent word embedding technologies inspire us to automatically detect concept drift in large-scale corpora [9, 6, 5] . However, comparing embeddings generated from different corpora is a complex task [6, 5]. How to visually inspect concept drift is also a challenge [16].

In this paper, we propose to use a simple approach to quantify the concept drift based on their contexts generated from two time periods and apply it to subsequent time periods to study concept drift over a long period of time. We study more than 50 thousand English books published between 1800 and 1950.

We first divide the whole period into subsequent 20-year time spans. Based on word embeddings corresponding to each individual time span, we calculate the context of each word at that particular period. Secondly, we measure the concept drift by comparing the contexts of the same word from different periods. Looking how the context changes from the starting period to the ending period, we can easily identify the most dynamic words over the 150 years. For these words which potentially underwent drastic change in their meaning, we further measure how their contexts change in subsequent time periods and visualise the change over time. Some interesting associations to historical events are also discovered.

2 Related Work

Diachronic semantic shifts has gained much attention because of the availability of large corpora and recent success in computing *distributional semantics* [4] in natural languages [12]. When computing distributional semantics, words are represented by sparse or, more recently, dense vectors based on their co-occurrences in a corpus. In other words, words are embedded in a semantic space and, more importantly, similar words are embedded nearby each other in this space.

To study semantic shift, it is therefore natural to first construct word embeddings in separated time periods before comparing these embeddings across time. However, the similarity between the word embeddings generated from separate time-specific corpora cannot be computed directly because the stochastic embedding algorithms could only roughly guarantee the stability of the pairwise similarities between words but the numerical embeddings are often rotated after each training, i.e. invariant under rotation. An earlier proposal was to incrementally update diachronic embedding models [10], where the word embeddings trained on the previous time period are used to initialise the training for the current period. Later researchers proposed to align these spaces by unifying the coordinates via local word alignments [11] or by projecting one space to another via orthogonial procrustes [6].

Another approach is to study the neighbours or the context of a word at two different time periods to measure semantic shift. Azarbonyad et. al. in [1] used the neighbors of a word to determine its stability. Their best model uses the traditional alignment-based method weighting the neighbors' rank and their stability, requiring computation on whole vocabulary. Later, Gonen et al. [5] simply took the top-k neighbors in each of the two corpora and measure the overlap of these two lists. A smaller overlap suggests more drastic change. They applied this method to corpora separated based on different criteria, such as age, gender, profession, time of tweet. However, the concept stability of neighboring words is unsure for a long period, for example over centuries. In this paper, we apply this method to study English literature that spans over 150 years.

More recently, researchers also proposed dynamic word embeddings models to jointly learn word embeddings across different times periods [2, 17]. By enforcing the alignments simultaneously, there is no need to train separate time-specific

embeddings, i.e., the resulting word embeddings are time-aware already. We will explore this aproach in the future.

3 Method

Here, we first describe the method which measures the changes in word meaning based on its contexts at two different time periods. Secondly, we divide the whole corpus into subsets corresponding to consequent time periods and study how the meaning of a word change over time.

3.1 Measuring drift based on context

Inspired by the method proposed in [5], given two periods in time, we collect separate corpora corresponding to each period. As shown in Figure 1, we generate embeddings for each word in the separate corpus. For each word, we calculate its context as the top K most similar words. For the same word that occurs in both corpora, we measure the similarity of its contexts at two different time periods. This similarity reflects the changes in word meaning: the more similar the two contexts are, the less change in meaning the word has.

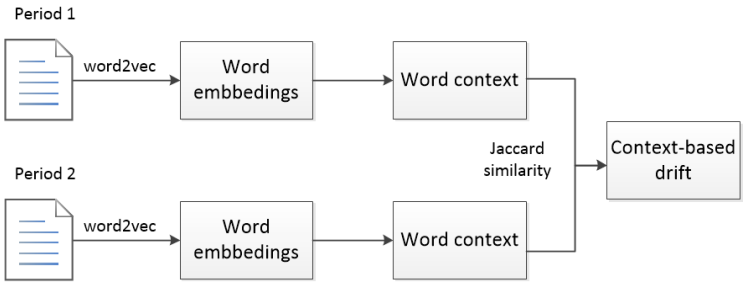


Fig. 1. workflow for measuring concept drift

Word embedding First, we need to generate embeddings for each word that occurs in each corpus. All words in the corpus are embedded as high-dimensional vectors, and semantically similar words embedded near to each other in a semantic space. The purpose of this step is to use numeric vectors to represent the meaning of words, so that the similarity between words could be computed easily.

Word context After the embedding spaces, two semantic spaces are generated from each corpus. These two semantic spaces can be aligned to use the same coordinate axes before we could compares words in these two spaces, as proposed in [6]. Here, we adopt the method proposed in [5] to use the closest neighbours of a word, i.e. its context, to reflect the extensional meaning of the word, therefore, the change in context at two different times reflects the drift in the meaning of the word. For each word, we select the top K most similar words as its context.

Drift based on context similarity For a word that occurs in both time periods, how much its context changes from one period to the other reflects the change in its meaning. Since the context is defined as the set of top K most similar words, we use Jaccard similarity coefficient [8] to measure the similarity between two contexts of the same word but from two different time periods. Jaccard similarity coefficient is a statistic used for gauging the similarity and diversity of sample sets. Let A and B are two sets, the Jaccard similarity coefficient is calculated as follows,

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} = \frac{\mid A \cap B \mid}{\mid A \mid + \mid B \mid - \mid A \cap B \mid} \tag{1}$$

A high Jaccard similarity coefficient suggests that the context of the word has not changed much, while a low value suggests that the meaning of the word might have changed over time.

3.2 Analysing concept drift over time

Given a corpus which spans a long period of time, it is then possible to study how individual words change over time. We divide the whole corpus into multiple subsets corresponding to subsequent time periods. We measure the drift of words from the beginning period to the last. This way, we could detect the most dynamic and static words over the long period of time. It is also possible to look more carefully when a word has undergone a critical moment when its meaning changed dramatically.

4 Data set

We download the full text content of 51,625 English books from Project Gutenberg.[1] Unfortunately, the exact years of publication for these books are missing. However, the birth and death years of the author are available in the data. We therefore took the average of the birth and death year of the author as the approximate year of publication for the book.

After grouping books by their year of publication, we find that, although the earliest books were written before 500 BC, the number of books published

[1] https://www.gutenberg.org/

earlier than the 18th century is far less than that from the 18th century to the mid 20th century. Because of the copyright restriction, books in recent years are also limited. In this study we focus on the books published from 1800 to 1950. We further divide the corpus into consequent groups of 20-year time spans, as hown in Table 1.

Group	Time span	Book
1800	1790–1810	1230
1830	1820–1840	2214
1850	1840–1860	3712
1870	1860–1880	6595
1890	1880–1900	8661
1910	1900–1920	7721
1930	1920–1940	1734
1950	1940–1960	1926

Table 1. Number of Books in each period

5 Experiment and results

5.1 Word embedding and context computation

For each time period show in Table 1, we trained a word2vec model using the gensim library[2] using the full text content of the books published within that period of time. We chose the continuous Bag-of-Words(CBOW) model [13], set the vector size as 100, the minimum count 10 (ignoring words that occur less than 10 times), the window size 20 (taking 20 words behind and 20 words ahead as the training context) and took the rest parameters as their default values. After embedding, each word at each time period is represented as a 100-dimensional vector. For each word at a particular time period, we calculated the top 20 most similar words as its context in that period.

5.2 Measuring drift

After the context of a word is calculated for each time period, we can now measure how much this word has drifted from one time period to another.

Sensibility of parameter K The model is generally stable. We examine how k parameter, which defines the length of similar words lists, affect the stability. We change k as [20,30,60,100], and the curves highly coincide with each other in the whole plot for Fig.2, so in most part of these plot, we regard the same

[2] https://radimrehurek.com/gensim/models/word2vec.html

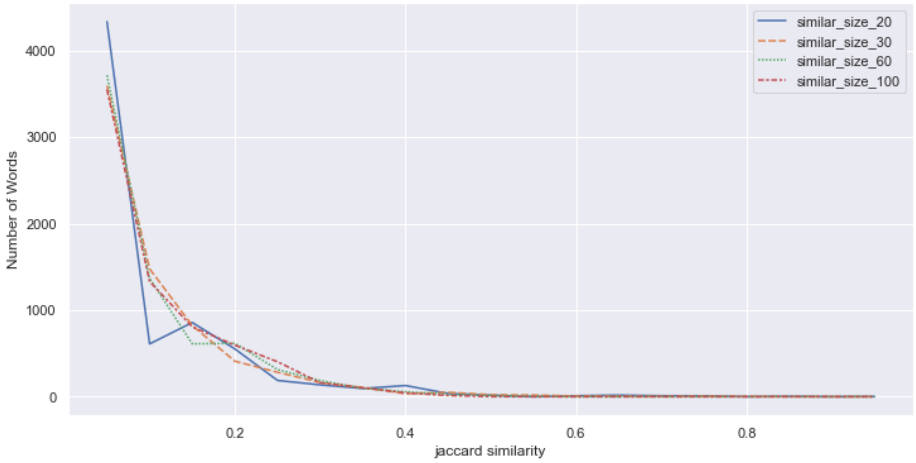


Fig. 2. Jaccard similarity in different K (length of similar words lists)

distribution as proof of the stability of our model. In the following experiments, we take k = 20.

We calculate the Jaccard similarity between the context at 1800 and that at 1950. The distribution of the Jaccard similarity is shown in Fig. 3. As the distribution shows, very few words have stable context throughout these 150 years. Words such as sister, daughter, mother, wife and husband are rather stable word, while other words such as witch, foster, hive and potion have changed dramatically.

Many words have completely changed their contexts. However, these words are mostly infrequent words, such as chestnut, hive, vantage, and coffin. Their embeddings and consequent contexts are over sensitive to the corpus. We could not make solid conclusions in terms of the drift of their meaning.

This still helps us to identify interesting cases of concept drift among the words that have a low context similarity. Once identified, we can dive into the more granule time periods and inspect the drift more closely. For example, Fig. 4 shows the drift of word peace over time. The Jaccard similarities between the current context and that of the previous period are plotted. The sharp decrease from 1890 to 1930 suggests that there was a drastic change of the meaning of the word peace in that period.

5.3 Visualising individual drift

The change in Jaccard similarity as shown in Fig. 4 only provides a signal of drift, but not the content of drift. To dive deeper into what exactly happened for specific words in specific periods, we present a visualization method that helps to see how words have changed.

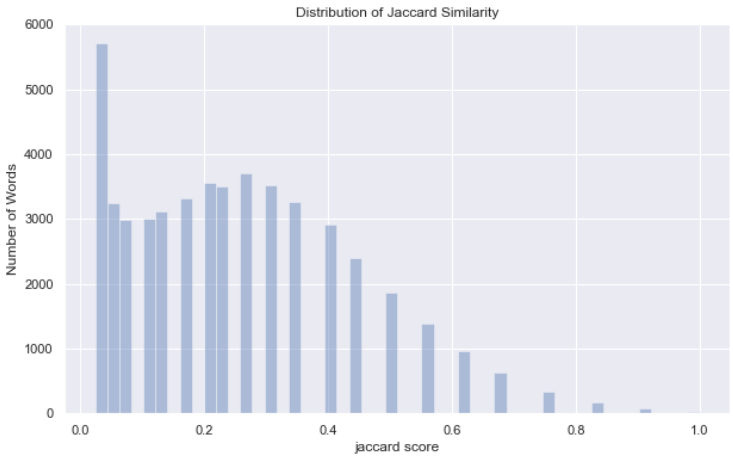


Fig. 3. The distribution of Jaccard Similarity between starting(1800) and ending(1950) period

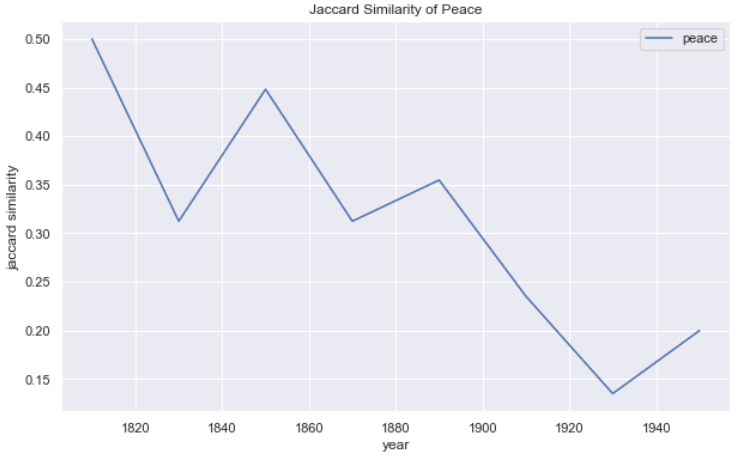


Fig. 4. Jaccard Similarity of 'peace' over time (comparing to the previous neighbor time group)

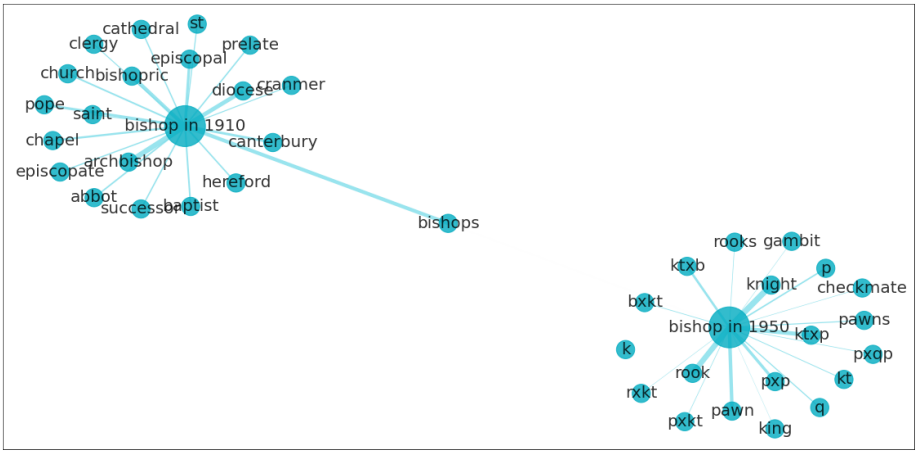


Fig. 5. The drift for bishop between 1800s and 1950s

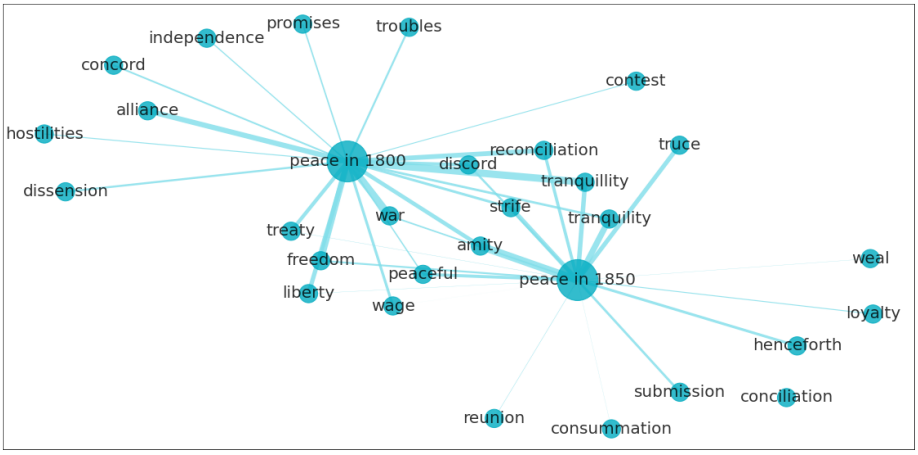
Our visualization is inspired by the work of Wijaya and Yeniterzi[16]. In their visualization, each word is a node and there is an edge between two words if they co-occur in the same context. The width of the edges is the frequency of co-occurrence. As shown in Fig. 5, our visualization consists of two clusters. One is the target word with its top 20 context words at the first time span. The other is the target word with its top 20 context words at the second time span. The line width shows the cosine similarity between the target word and the context word.

In Fig. 5, the word bishop at 1910 is mostly associated with religious words, such as prelate, church, cathedral, and dioce while at 1950, it is more related to the chess game, as its context includes words like checkmate, pawn, and knight. It is worth mentioning that ktxb, bxkt, and rxkt are notation words in chess but not meaningless words.

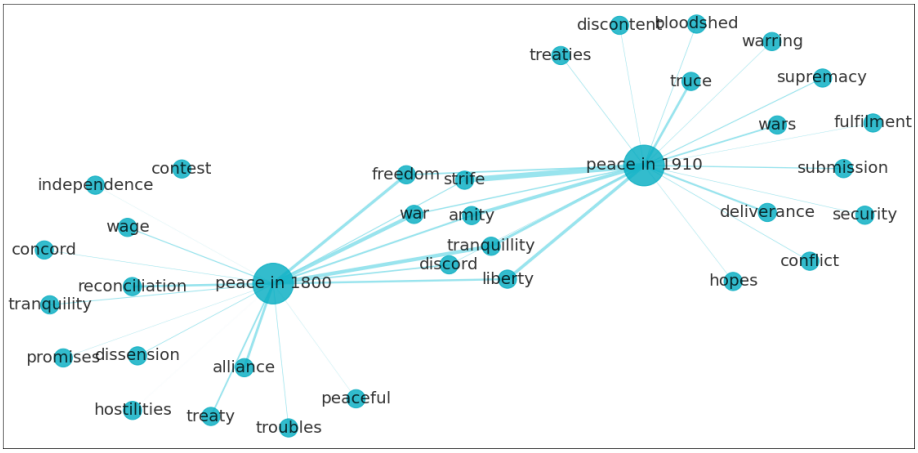
Compared to the popular t-SNE visualization [14] used in [5], our visualization is more comprehensible and intuitive to compare the intersection and the unique section. It can also be used for tracking how the concept of a word changed gradually over a long time span. However, this visualization often become unreadable because of the complexity when the number of the context words increases.

Fig.6 shows the sequential change of the word peace from 1800 to 1950. Fig.6 (b) shows that war is in the same context with peace at 1800 and 1910. It suggests that peace and war were often mentioned together .

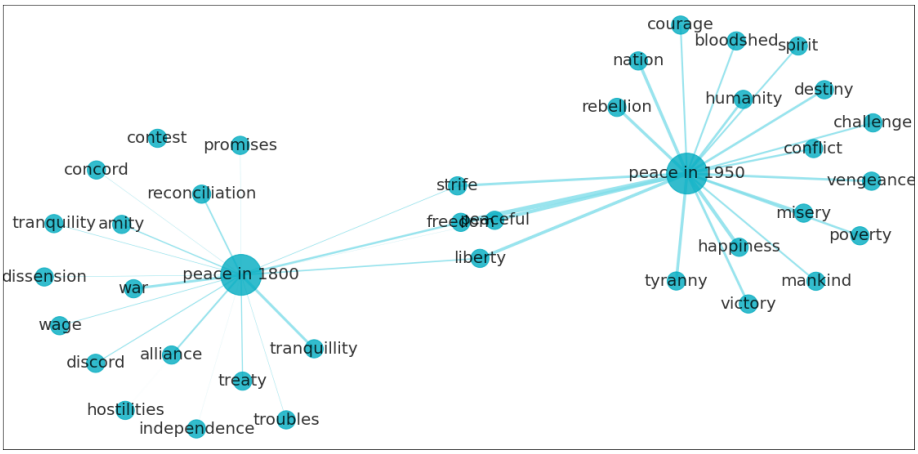
However, the link between war and peace disappeared in 1950. New relationships emerged, such as spirit, humanity, tyranny, and poverty. Link to the statistics of war. There were high-intensity conflicts around the 1800s (Napoleonic Wars, etc), 1860s (American Civil War, etc.), and 1910s (World War I). There were few wars after World War II (1945). It makes sense that people transferred their



(a) The word peace at 1800 and 1850



(b) The word peace at 1800 and 1910



(c)The word peace at 1800 and 1950

Fig. 6. Concept Drift of the word peace over time

concerns about peace into other topics like spirit, humanity, tyranny, and poverty in the 1950s.

Because of the copyright restriction, Gutenberg data set only has few books after 1970. It limits us to apply our method to recently published books. A potential future work for this work is to apply this approach on possible contemporary book corpus. There would be more interesting findings closer to our life.

6 conclusion

Concept drift reflects the change of the social consensus. Detecting word usage in different periods is an important research method. We propose a computational approach to discover the drastic concept drifts by their context in the historical English books over centuries. It quantifies the extend of concept drift and makes the rank of drastic change possible. We also present a new way to compare the concept of a word in different periods. We show that our visualization is simple and intuitive. It also has the unique advantage of demonstrating the gradual change of concept overtime.

References

1. Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1509–1518, New York, NY, USA, 2017. Association for Computing Machinery.
2. Robert Bamler and Stephan Mandt. Dynamic word embeddings. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.
3. Oxford English Dictionary and WALTON Street. Oxford english dictionary. *Retrieved February*, 4:2019, 2019.
4. John Firth. *A synopsis of linguistic theory. 1930-1955.* Blackwell, 1957.
5. Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, 2020.
6. William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.
7. Archie Hobson. *The Oxford dictionary of difficult words.* Oxford University Press, USA, 2004.
8. Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

9. Adam Jatowt and Kevin Duh. A framework for analyzing semantic change of words across time. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 229–238. IEEE, 2014.
10. Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.
11. Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.
12. Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
13. Tomas Mikolov, Kai Chen, G. S. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.
14. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
15. Shenghui Wang, Stefan Schlobach, and Michel Klein. Concept drift and how to identify it. *Journal of Web Semantics*, 9(3):247–265, 2011. Semantic Web Dynamics Semantic Web Challenge, 2010.
16. Derry Tanti Wijaya and Reyyan Yeniterzi. 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, 2011.
17. Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 673–681, New York, NY, USA, 2018. Association for Computing Machinery.