Tracing Shifting Conceptual Vocabularies Through Time

Gabriel Recchia[*], Ewan Jones[*], Paul Nulty[*], John Regan[*], and Peter de Bolla[*]

[*]The Concept Lab, CRASSH, University of Cambridge, Cambridge, United Kingdom
{glr29,ejj25,pgn26,jjr35,pld20}@cam.ac.uk

Abstract. This paper presents work in progress on an algorithm to track and identify changes in the vocabulary used to describe particular concepts over time, with emphasis on treating concepts as distinct from changes in word meaning. We apply the algorithm to word vectors generated from Google Books n-grams from 1800-1990 and evaluate the induced networks with respect to their *flexibility* (robustness to changes in vocabulary) and *stability* (they should not leap from topic to topic). Finally, we describe work in progress using the British National Biography Linked Open Data Serials to construct a "ground truth" evaluation dataset for algorithms which aim to detect shifts in the vocabulary used to describe concepts.

Keywords: concepts · word embeddings · Linked Open Data

1 Introduction

Some influential theories of conceptual structure, such as the so-called *name priority view* [1] and some interpretations of the *classical theory of concepts* [2], treat concepts[1] as essentially in one-to-one correspondence to word senses [3,4,5]. On this view, one word might have several different senses and thereby correspond to several different concepts, but it is nonetheless possible to identify concepts via a careful examination of word meanings. Some modern philosophers and psychologists have made convincing arguments that this view is overly simplistic or flat-out wrong [1,6]. Even if one does believe in a direct correspondence between word senses and concepts, however, it is clear that a change in word sense does not necessarily entail a change in the concept that was originally associated with it. For example, the word

[1] Rather than thinking of concepts in a way that strongly links them to a particular lexeme (e.g., "the concept of justice"), we have argued elsewhere that it is preferable to think of concepts (at least insofar as they are expressed in discourse) in terms of their functions, one of which is to permit two interlocutors to sense that they have arrived at a common understanding of the matter under discussion. This is rather different and more abstract than the notion of a concept as being equivalent to a class in a classical ontology, and more specific than a theme or topic. However, for purposes of clarity and compatibility with the way related work speaks about "concepts," our use of the word in this paper roughly conforms to the vague OED definition of "a general idea or notion." We are explicitly *not* using it to refer to "the meaning that is realized by a word or expression."

broadcast started to change from having the meaning of “scattering [seed] abroad over the whole surface, instead of being sown in drills or rows” to being associated with the transmission of radio or television signals in the 1920s [7,8]. However, the fact that the primary sense of *broadcast* changed did not mean that the concept of sowing seeds over a wide area went away. Similarly, it seems clear that a culture could possess a particular concept even if no corresponding word or collocation exists in the primary language spoken by members of that culture.

The distinction between word senses and concepts is an important one to draw because, as pointed out by Wevers et al. [9], some computational approaches described as methods for detecting changes in “concepts” are often actually methods for detecting changes in the use of a single word or an unchanging group of words over time. Because word senses change over time, a change in the frequency or lexical associations of a particular word does not necessarily entail a change in the concept of interest. Being able to track concepts over time in a way that is robust to shifting vocabularies is therefore essential.

Methods for detecting conceptual change in time-varying textual sources are particularly relevant in the context of Linked Open Data (LOD). To assist in the maintenance of LOD ontologies, knowledge engineers may wish to use time-varying text corpora, such as academic journals or news sources, to monitor conceptual change over time. Consider someone who maintains an ontology intended to represent relationships between various concepts in the neuroscience literature, who notices that this year there has been a marked uptick in the frequency of particular words that previously occurred only rarely. Does this merit the addition of a new class to the ontology? Or is this simply novel language for describing an old idea? Ultimately, this must come down to human judgment, but automatic methods for assisting with the decision could highlight important related classes already represented in the knowledgebase.

This paper presents work in progress toward an algorithm to track vocabulary associated with particular concepts over time in a flexible and stable way. In the next section, we describe related work, particularly a promising model recently developed by [10]. In Section 3 we implement a model which avoids one of the weaknesses of previous work while retaining the most important benefits. Finally, we describe work in progress using the *British National Biography Linked Open Data Serials* to construct a “ground truth” evaluation dataset for algorithms of this sort.

2 Related Work

To address the problem of word sense change described in the Introduction, the *Concepts Through Time* model advocates an alternative approach to tracking concepts, using a set of Dutch newspapers from 1890-1990 as a corpus [9]. Rather than selecting a static set of terms and monitoring its frequency over the entire century, they select an initial term or terms of interest and find a cluster of words that are highly similar, according to a word embedding model trained on a specific timeslice (e.g., articles from the years 1890-1900). The cluster is updated from timeslice to subsequent timeslice in a manner which acknowledges that "the set of words used to discuss a particular concept might not show any overlap at all between different periods of time" [9]. However, treating time-shifted collections of words with no overlap whatsoever as the 'same' has its own drawbacks. As [11] points out, "Imagine a subset of documents containing strong co-occurrence patterns across time: first between birds and aerodynamics, then aerodynamics and heat, then heat and quantum mechanics—this could lead to a single topic that follows this trajectory, and lead the user to inappropriately conclude that birds and quantum mechanics are time-shifted versions of the same topic." Perhaps for this reason, subsequent work by the developers of *Concepts Through Time* notes that "a successful system… should strike a balance between an adaptive strategy that responds to changes in vocabulary, and a more conservative approach that keeps the vocabulary stable" [10]. Their revised model requires a user to select an initial set of seed terms and an algorithm to construct vocabularies of related terms for each timeslice: *adaptive, nonadaptive,* or *hybrid*. The adaptive method is most relevant to the present work. This method starts with an input vocabulary (initially the user's set of seed terms), expands that by adding words exceeding some minimum similarity threshold to the set, constructs a network from this set such that all pairs of nodes (words) exceeding the threshold are assigned an edge, and then prunes nodes that are low in degree centrality. The resulting words are used as the input vocabulary for the next timeslice, and the process repeats until the final timeslice.

Topic models have been another popular approach for monitoring groups of related words over time [11,12,13,14]. However, these often either do not explicitly model changes in vocabulary within a particular topic/concept, or do not pay explicit attention whether the method allows topics to drift far afield from their original conceptual content. One contribution of the present work is that it does both, while resolving an important difficulty with the most similar approach we are aware of.

Finally, much work has been done on automatically tracing changes in a given word's meaning over time, e.g. [8,15,16]. Although this clearly differs from our aim of tracing changes in the vocabulary used to describe particular concepts, these methods are extremely useful for our purposes. For example, a word may need to be excluded from a core of tightly interrelated terms if its meaning drifts too far afield from the rest. We therefore here make extensive use of the *HistWords* vectors [8] developed by applying skip-grams with negative sampling (one of the algorithms available in *word2vec*) to n-grams distributed by Google Books. *HistWords* contains a separate vector for each of a very large number of terms for every decade from 1800 to 1990, such that words that appear in similar contexts within a given timeslice have similar vectors. Such vectors successfully capture shifts in word meaning over time [8], and we use the same approach and data to quantify semantic similarity. We describe how we use these vectors in more detail in the following section.

3 Time-Varying Relationships in Text

Recall that the adaptive method of [10] involves an expansion step in which words related to any word in the input vocabulary are added to the network as nodes, and a pruning step in which nodes low in network centrality (in-degree or out-degree) are pruned. Although this is an excellent way to pull in novel vocabulary while also preventing the overall network from drifting too far afield, it has one unintended consequence. When the input vocabulary contains a word linked to two densely connected but unrelated clusters (e.g., a polysemous word), unrelated clusters of words will be added during the expansion step (Figure 1). Because nodes in each cluster have high degree, they will not be eliminated in the pruning step. The consequence is that unrelated, weakly connected clusters can persist as part of the same "concept." The example in Figure 1 makes this particularly clear by illustrating two clusters so unrelated that they would become disconnected if the node connecting them were pruned, but it is important to recognize that this phenomenon remains a problem even if a constraint were imposed requiring the graph to be fully connected. The other two methods described by [10] (*nonadaptive* and *hybrid*) suffer from the same difficulty.

Our method addresses this by allowing two nodes to be treated as part of the same "conceptual network" only if *all* words in the network are highly related to *all* other words in the network. Because we first describe how this method can be used to track concepts in diachronic text corpora, we use "nodes" and "words" interchangeably, and "relatedness," "similarity," and "edge weight" as synonymous with "cosine similarity" (e.g., similarity between word vectors in the *HistWords* data). Like [10], we treat documents from every timeslice (in our case, decades from 1800-1990) as a separate subcorpus and build a separate vector space corresponding to each.

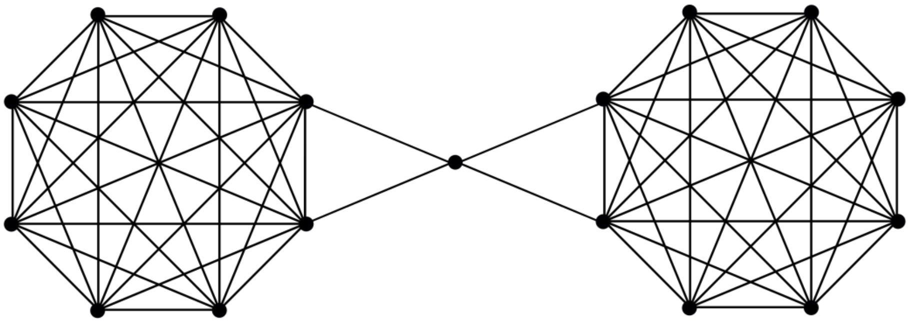

Fig. 1. A graph representing two unrelated clusters of words connected by a single polysemous word. Even if the center node is eliminated in the pruning step described in [10], the nodes in each 8-clique may not be, due to their high degree. Both clusters thus erroneously continue to be interpreted as part of the same "concept."

3.1 Algorithm

Given a size *k* and a seed set of words W, the algorithm begins by finding the fully connected graph of size *k* containing all words in W such that the minimum edge weight (in the earliest timeslice) is as high as possible. This can be done efficiently by attempting to find a subgraph of size *k* containing all words in W such that every edge exceeds a very high threshold[2], but then gradually lowering the threshold until an appropriate subgraph is found. Afterwards, the vectors for the second timeslice are loaded, and the subgraph is updated by attempting to answer the question, "Is it possible to increase the minimum edge weight by replacing one of these nodes with a node currently not in the subgraph? If so, which of all possible replacements would increase the minimum edge weight the most?[3]" Because typically only one edge is equal to the current minimum edge weight, this can also be computed efficiently. This corresponds to a "drop one, add one" rule where, for any given timeslice, a single word from the network of the previous timeslice will be replaced if and only if doing so increases the minimum similarity between every word pair in the resulting network. The process repeats for every subsequent timeslice. Table 1 illustrates an example of an evolving network built using this method.

Our primary concerns were that conceptual networks be traced in such a way that is *flexible* (words whose meanings shift away from the conceptual core should drop out) but also *stable* (a network initially about birds should not drift to quantum mechanics). We tested the model by initializing it with 500 words randomly selected from the 30,000 most frequent terms in *HistWords*, which were used as the lexicon. Of these, there were 212 words such that a fully connected network of size 9 existed with a minimum edge weight of 0.2 or greater could be constructed.

1900	affable,cheerful,courteous,gay,genial,humored,natured,sprightly,witty
1910	affable,cheerful,courteous,gay,genial,humored,natured,humoured,witty
1920	affable,cheerful,courteous,gay,genial,jovial,natured,humoured,witty
1930	affable,cheerful,courteous,gay,mannered,jovial,natured,humoured,witty
1940	affable,cheerful,courteous,amiable,mannered,jovial,natured,humoured,witty
1950	affable,cheerful,courteous,amiable,mannered,vivacious,natured,humoured,witty
1960	affable,cheerful,courteous,amiable,mannered,charming,natured,humoured,witty
1970	affable,cheerful,courteous,amiable,mannered,charming,natured,gentle,witty
1980	affable,cheerful,courteous,amiable,humored,charming,natured,gentle,witty
1990	affable,cheerful,courteous,amiable,clever,charming,natured,gentle,witty

Table 1. Evolution over time of the network constructed from the seed word "gay."

A potential criticism of this model is that while it purports to be 'flexible' in the sense that it traces a group of *conceptually related* words (rather than merely words associ-

[2] Because the threshold is initially set so high that no such subgraph can be found, this method ensures that the first subgraph discovered which meets these criteria is the one desired.

[3] Note that every node in the subgraph must correspond to a unique word.

ated with the seed term alone), the fact that it is initialized with words closely related to the seed term may mean that in practice the seed term always ends up as a permanent part of the network. Table 1 illustrates that in at least one case of radical semantic change (the word “gay”), the seed term does successfully drop out by the 1940s. However, it is possible that this virtually never happens. Flexibility was therefore evaluated by quantifying the proportion of the 212 initial networks (year 1800) in which the seed term did drop out by 1990. Another potential criticism is the reverse: Because every timeslice offers an opportunity to jettison one term and incorporate a new one, networks might drift to completely different topics. For example, if each iteration caused a random word to be replaced with a word not previously in the network, then after 19 timesteps a typical graph of size 9 would be expected to retain only $(8/9)^{19}$ = 10.7% of its initial vocabulary. We therefore also computed the proportion of the vocabulary shared in the 1800 vs. 1990 clusters, with qualitative analysis of the clusters with the least shared vocabulary, to evaluate whether they exhibited less drift than this random baseline.

3.2 Results

With respect to flexibility, the seed word used to generate the initial size-9 network in 1800 was no longer present in the 1990 network in 147 of 212 cases (69%). In 91% of these cases, the seed word never re-entered the network once it had dropped out, suggesting that the seed word can indeed be permanently ejected from the conceptual core if its meaning or associations drift in a different direction. With respect to stability, the average overlap in vocabulary between the initial 1800s network and the final 1990s network was 33%, with all 212 cases sharing at least one word (11%) in common with the original 9-word network. Even when only one word was shared, the network typically did not drift too far afield, as in the case of the seed word “uneasy” (1800: *anxieties, dejected, dejection, distraction, fits, insupportable, languishing, uneasy, weariness*; 1990: *anxieties, grief, despair, disappointment, misery, sorrow, anguish, sadness, loneliness*). The full set of networks generated in this evaluation may be obtained from http://nowin2d.com/vocabularies.html.

3.3 Discussion

The results suggest that even with such a rigid algorithm, the induced conceptual vocabularies are certainly flexible and reasonably resistant to drift. However, there were occasional clear cases in which vocabulary drifted significantly away from the original conceptual core, as in the case of the network generated from the seed word “logical” (1800: *abstruse, definitions, disquisition, disquisitions, explanations, explication, grammatical, illustrating, logical*), which drifted towards different areas of academic study by the 1990s (*abstruse, mathematical, philosophy, theory, metaphysics, metaphysical, empirical, theoretical, philosophical*). One obvious direction for future work is the optimization of initialization parameters (network size, initial edge weight threshold), which were chosen arbitrarily. However, the fact that even arbitrarily chosen initialization parameters resulted in reasonably flexible and stable networks is promising. Other directions could include making use of the information about how a conceptual vocabulary has changed in the past to predict how it will change in the future, using techniques that parallel those applied in predicting ontology evolution, e.g. [17]. In addition, future work should consider a stronger evaluation metric. The final section describes work in progress towards constructing a "ground truth" evaluation dataset that could be used to do just that.

4 Constructing ground truth evaluation data from LOD

To more fully evaluate an algorithm's ability to track vocabulary change associated with arbitrary concepts, a "ground truth" dataset is necessary. The only such data of which we are aware are offered by [10]. However, this is limited to 21 concepts spanning only four decades and is in Dutch, making it incompatible with most large, diachronic corpora. However, something very near to a much larger, English-language ground truth dataset already exists as LOD, in the form of the British National Bibliography Linked Open Data (BNBLOD) collections. Although we see 'concepts' as nonidentical to subjects as defined by the BNB, it is nonetheless likely that there is a high level of conceptual relatedness between all documents to which the British Library has assigned the subject *http://bnb.data.bl.uk/id/concept/lcsh/Engineering*, even if the vocabulary of such documents differs markedly from year to year. Particularly useful are the BNBLOD Serials, which in addition to including the year of each journal's first publication, very commonly contain "Journal of X" in the title, where "X" corresponds to a short phrase describing a particular subject. The vocabulary of such 'title phrases' is often tied to a particular moment in time. Consider, for example, the phrases so extracted from the earliest "journal of X" journals in the BNBLOD Serials assigned the subject of *Psychiatry* (1876: 'nervous and mental disease'), *Engineering* (1921: 'applied mathematics and mechanics'), *Entrepreneurship* (1985: 'business venturing'), and *Tourism* (1972: 'travel research'). These phrases are no longer commonly used to describe these subjects. As a first step in constructing an evaluation dataset, therefore, we are first simply extracting title phrases, publication dates, and subjects from the BNBLOD Serials and structuring them as follows: Given a start year y_1, an end year y_2, and a phrase extracted from the title of a serial having subject S first published in y_1, the algorithm being evaluated must predict which words and phrases are most likely to appear in titles of other journals of subject S which were published in y_2. A robust algorithm trained on an appropriate could ideally correctly identify that the cluster of words that contains, e.g., "business venturing" in 1985 ought to include "entrepreneurship" by 1995 (rather than, say, "business organization"), and that "travel research" in 1972 is closer to "tourism" in 1992 than to "educational travel". It should be noted that this is just a first step, and we hope to include other methods of evaluation with time. It is our hope that such a dataset will allow us not only to better evaluate our own research but move the field of representing diachronic conceptual change forward as a whole.

5 Acknowledgments

This work was supported by a private donation to the Cambridge Centre for Digital Knowledge (CCDK) at the University of Cambridge.

6 References

1. Seiler, T.B., Wannenmacher, W. (eds.): Concept Development and the Development of Word Meaning (Vol. 12). Springer Science & Business Media, Berlin (2012)
2. Margolis, E., Laurence, S.: Concepts. In: E. N. Zalta (ed.), The Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/archives/spr2014/entries/concepts/ (2014)
3. Fodor, J.A.: The Language of Thought. Crowell, New York (1975)
4. Clark, E.V.: Meaning and Concepts. In: P. H. Mussen (ed.), Handbook of Child Psychology, vol. 3: Cognitive Development, pp. 787–840. Wiley, New York (1983)
5. Murphy, G.: The Big Book of Concepts. MIT Press, Cambridge (2002)
6. Glanzberg, M.: Meaning, Concepts, and the Lexicon. Croatian Journal of Philosophy 11(1), 1-29 (2011)
7. OED Online.: "Broadcast". Oxford University Press, http://www.oed.com
8. Hamilton, W. L., Leskovec, J., Jurafsky, D.: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. arXiv preprint arXiv:1605.09096 (2016)
9. Wevers, M., Kenter, T., Huijnen, P.: Concepts Through Time: Tracing Concepts in Dutch Newspaper Discourse (1890-1990) Using Word Embeddings. In: Digital Humanities 2015, Sydney (2015)
10. Kenter, T., Wevers, M., Huijnen, P.: Ad Hoc Monitoring Of Vocabulary Shifts Over Time. In: Proc. 24th ACM International on Conference on Information and Knowledge Management (pp. 1191-1200). ACM, New York (2015)
11. Wang, X., McCallum, A.: Topics Over Time: A Non-Markov Continuous-Time Model Of Topical Trends. In: Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 424-433). ACM, New York (2006)
12. Blei, D.M., Lafferty, J.D.: Dynamic Topic Models. In: Proc. 23rd International Conference on Machine Learning (pp. 113-120). (2006)
13. Hall, D., Jurafsky, D., Manning, C.D.: Studying The History of Ideas Using Topic Models. In: Proc. Conference on Empirical Methods on Natural Language Processing (EMNLP) (pp. 363-371). Association for Computational Linguistics, East Stroudsburg, Pennsylvania. (2008)
14. Sigrist, R., Rawat, V.: Topic Evolution In A Stream Of Documents. In: Proc. SIAM International Conference on Data Mining. SIAM, Philadelphia (2009).
15. Gulordava, K., Baroni, M.: A Distributional Similarity Approach To The Detection of Semantic Change in the Google Books N-gram Corpus. In: Proc. of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS) Workshop. Association for Computational Linguistics, East Stroudsburg, Pennsylvania. (2011)
16. Wijaya, D.T., Yeniterzi, R. Understanding Semantic Change Of Words Over Centuries. In: Proc. DETECT (International Workshop on DETecting and Exploiting Cultural diversiTy on the social web) (pp. 35-40). ACM, New York (2011)
17. C. Pesquita, F. M. Couto.: Predicting the extension of biomedical ontologies. In: PLoS Computational Biology, 8(9):e1002630. (2012)