Diachronic and diatopic word embeddings from British historical newspapers

ABSTRACT

This poster presents a new resource for the study of diatopic semantic variation in historical texts consisting of word embedding models trained on historical British newspapers for five geographical regions. We discuss the embedding models and present an analysis on the lexicon of mechanisation in 19th-century English. As an application of our models, we show how different results were obtained from running a changepoint detection algorithm on the embeddings for the North and South of England, historically corresponding to a major socio-political split in Britain. This suggests how the semantic change of words related to the mechanisation process following the Industrial Revolution did not occur at the same pace across British regions. Our methods can be applied to other languages and historical texts, and our resources can be reused to investigate other questions related to semantic change in 19th-century English.

KEYWORDS

Historical corpora, word embeddings, semantic variation, semantic change.

1. BACKGROUND AND OBJECTIVES

Interpreting meaning expressed in text is a fundamental aspect of humanistic research. With the recent growth in the availability of historical texts in digital format, researchers now have the opportunity to mine these collections at scale. Computational methods allow us to conduct a variety of semantic analyses on large textual collections, including detecting evolving word meanings over time [20] and tracing conceptual change [4], which can support research on socio-cultural phenomena (e.g. [11]). State-of-the-art techniques rely on word embeddings to generate low-dimensional vector representations of words from words' co-occurrence data [8, 16] which capture important semantic properties of words, including similarity and analogy relations. Recent years have witnessed a growth in the number of research projects generating diachronic word2vec embeddings [14] from historical texts. Researchers have released word embeddings trained on various diachronic corpora [5, 6, 7, 10]. Diatopic variation has been the object of much research in variational linguistics, with most studies focussed on synchronic data (e.g. [3, 18, 21], but less attention has been devoted to its quantitative study in diachronic contexts. This poster presents a new resource for the study of diatopic semantic variation in historical texts consisting of word embedding models trained on historical British newspapers for five different geographical regions, building on the experiment in [15] by adding the diatopic dimension to the diachronic one. The objectives of this study are to discuss the embedding models and to outline the results of our analysis on the lexicon related to mechanisation in 19th-century English. Our methods are general enough to be relevant to research on other languages and historical texts.

2. DATA AND METHODS

We used a corpus of historical British newspapers comprising around 4.6 billion tokens and spanning the period between 1801 and 1920. The corpus includes titles specifically selected for the Living with Machines project[1] (2.3 billion tokens) and selected titles from the British Library's Heritage Made Digital digitization project (further 2.3 billion tokens).[2] Using the place and year of publication of each newspaper, we divided the corpus into two subcorpora containing articles published in two broad geographical regions, Northand South England, historically corresponding to one of the main socio-political divides in Britain,and we split each geographical subcorpus into 10-year slices.[3] Because of the size of the corpus, the texts underwent minimal pre-processing (lowercasing, punctuation and stopword removal) and no lemmatization. We then trained Word2Vec [14, 17] models for each decade in each geographical subcorpus, which we release as a resource for the community.[4] We also use the embeddings to trace the semantics of lemmas related to mechanisation across different decades. We aligned the semantic spaces via Orthogonal Procrustes [19] and used the cosine similarity between vectors across different decades to measure their semantic shift and the pruned exact linear time (PELT) algorithm [9] to detect potential semantic changes in each geographic subcorpus. The results from the

[1] https://www.turing.ac.uk/research/research-projects/living-machines

[2] https://www.bl.uk/projects/heritage-made-digital

[3] Training of diachronic models for additional British regions (Midlands, Scotland, and Wales) is also underway thanks to additional historical newspaper data from the British Newspaper Archive.

[4] The models can be found at https://doi.org/10.5281/zenodo.7892460; the associated code to train diachronic word embeddings can be found at https://github.com/Living-with-machines/DiachronicEmb-BigHistData.

overlapping time-slices between the subcorpora[5] were compared against each other to assess whether their semantic shift occurred virtually simultaneously across the two regions or whether some degree of diatopic variation could be posited.

3. RESULTS

Word	North England	South England
bulbs	1860s	-
cars	1860s	-
machines	-	1860s
match	1860s	1840s
matches	1860s	-
stamp	1860s	1860s
stamps	-	1840s
stock	-	1860s
trade	1860s	-
trolley	*not in vocabulary*	1850s

Table 1. Changepoints detected in the North-England and South-England subcorpora.

Our preliminary results show differences between the changepoints detected by the PELT algorithm for the North and South of England. While in some cases, like *match* and *stamp*, a changepoint was detected for both regions, for other words, such as *machines* and *stock*, a changepoint was only detected for the South, whereas for others, like *trade*, *bulbs* and *cars*, only for the North. For some words, a changepoint was detected in a region either only in the singular or the plural form, as in *match* (but not *matches*) in the South or *stamp* (but not *stamps*) in the North of England; in other cases a changepoint was detected earlier for one of the two forms. These subtle differences may have to do with the different usages and therefore different triggers for semantic change associated with some words in the singular and plural, for example when they are used as plural generics [12, 13], as opposed to referring to multiple instances of a concrete object. Moreover, even for words with a potential changepoint in both regions, the decade in which the shift occurred may differ: for *match*, for instance, a changepoint was detected later in the century for the North than for the South. Existing corpus query tools such as BNClab [1] enable lexical analyses of language usage across time; however, they do not offer a semantic search functionality, therefore they do not allow us to look for evidence of such semantic shifts in diatopic variation. To interpret these results and evaluate whether they correspond to historically driven intuitions, we plan to extract the nearest neighbours of the words for which a changepoint was detected in any of the models in the decades before and after the changepoints. The words found among the nearest neighbours should help us identify the type of semantic change which occurred for a given word. The shift in nearest neighbours (to be expected given the detected changepoints) can also be compared across different regions, to check whether the same kind of shift occurred across the board or whether, besides the difference in changepoint, a further difference in the type of semantic change can be observed.

[5] The corpus used for this preliminary experiment is imbalanced in temporal coverage: articles from North England cover the span 1830s-1910s, whereas those from South England 1800s-1880s. Additional models covering the remaining decades for each region are also underway thanks to the British Newspaper Corpus. For the purpose of this poster, we only present results from the portion of the century for which both North and South England have some temporal coverage (i.e. 1830s-1880s).

ACKNOWLEDGEMENTS

Work for this paper was carried out as part of *Living with Machines.* This project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC grant AH/S01179X/1), with The Alan Turing Institute, the British Library and King's College London, East Anglia, Exeter, and Queen Mary University of London. Newspaper digitisation by the British Library, undertaken as part of the Living with Machines project, was funded by the AHRC.

REFERENCES

[1] Brezina, V., Gablasova, D. & Reichelt, S. (2018). “BNClab”. http://corpora.lancs.ac.uk/bnclab [electronic resource, last accessed May 2023], Lancaster University.

[2] Camacho-Collados, Jose and Mohammad Taher Pilehvar. “From word to sense embeddings: a survey on vector representations of meaning,” *Journal of Artificial Intelligence Research* 63, no. 1 (2018): 743–788.

[3] De Pascale, Stefano. Token-based vector space models as semantic control in lexical lectometry. KU Leuven: PhD thesis, 2019.

[4] Fokkens, Antske, Serge ter Braake, Isa Maks and Davide Ceolin. “On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change,” *Proceedings of Drift-a-LOD@EKAW* (2016).

[5] Grayson, Siobhán, Maria Mulvany, Karen Wade, Gerardine Meaney and Derek Greene. “Novel2Vec: Characterising 19th Century Fiction via Word Embeddings,” *Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science*, (2016): 20–21.

[6] Hamilton, William L., Jure Leskovec and Dan Jurafsky. (2016). “HistWords: Word Embeddings for Historical Text”. https://nlp.stanford.edu/projects/histwords/ [electronic resource, last accessed May 2023].

[7] Hosseini, Kasra, Kaspar Beelen, Giovanni Colavizza and Mariona Coll Ardanuy. “Neural Language Models for Nineteenth-Century English,” *Journal of Open Humanities Data* 7, no. 22 (2021). http://doi.org/10.5334/johd.48

[8] Joulin, Armand, Edouard Grave and Piotr Bojanowski Tomas Mikolov. "Bag of Tricks for Efficient Text Classification," *Proceedings of EACL 2017*, (2017): 427–431.

[9] Killick, Roberta, Paul Fearnhead and Idris A. Eckley. “Optimal detection of changepoints with a linear computational cost,” *Journal of the American Statistical Association* 107, no. 500 (2012): 1590–1598.

[10] Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde and Slav Petrov. “Temporal Analysis of Language through Neural Language Models,” *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, (2014): 61–65.

[11] Kozlowski, Austin C., Matt Taddy and James Evans. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings,” *American Sociological Review* 84, no. 5, (2019): 905–949. https://doi.org/10.1177/0003122419877135

[12] Leslie, Sarah J., Sangeet Khemlan, Sandeep Prasada and Sam Glucksberg. “Conceptual and linguistic distinctions between singular and plural generics,” *Proceedings of the 31st Annual Cognitive Science Society*, (2009): 479-484.

[13] Mari, Alda, Claire Beyssade and Fabio Del Prete. *Genericity*. Oxford: Oxford University Press, 2012.

[14] Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. “Efficient estimation of word representations in vector space,” *Proceedings of Workshop at the International Conference on Learning Representations*, (2013).

[15] Pedrazzini, Nilo and Barbara McGillivray. “Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers,” *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, (2022): 85–95.

[16] Pennington Jeffrey, Socher Richard, Manning Christopher D. “Glove: Global Vectors for Word Representation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014): 1532–43.

[17] Řehůřek, Radim and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora,” *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (2010): 45–50.

[18] Ruette, Tom, Dirk Speelman and Dirk Geeraerts. “Measuring the lexical distance between registers in national varieties of Dutch,” Proceedings of the International Conference on Pluricentric Languages, (2011): 541–554.

[19] Schönemann,Peter H. “A generalized solution of the orthogonal procrustes problem,” *Psychometrika* 31, (1966): 1–10.

[20] Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu and Simon Hengchen. (Eds.). *Computational approaches to semantic change* (Vol. 6). Language Science Press, 2021.

[21] Wieling, Martijn, Simonetta Montemagni, John Nerbonne and Harald R. Baayen. “Lexical Differences between Tuscan Dialects and Standard Italian: Accounting for Geographic and Sociodemographic Variation Using Generalized Additive Mixed Modeling,” *Language* 90, no. 3 (2014): 669–92.