Journal of Open Humanities Data

Neural Language Models for Nineteenth-Century English

DATA PAPER

KASRA HOSSEINI
KASPAR BEELEN
GIOVANNI COLAVIZZA
MARIONA COLL ARDANUY
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present four types of neural language models trained on a large historical dataset of books in English, published between 1760 and 1900, and comprised of ≈5.1 billion tokens. The language model architectures include word type embeddings (word2vec and fastText) and contextualized models (BERT and Flair). For each architecture, we trained a model instance using the whole dataset. Additionally, we trained separate instances on text published before 1850 for the type embeddings, and four instances considering different time slices for BERT. Our models have already been used in various downstream tasks where they consistently improved performance. In this paper, we describe how the models have been created and outline their reuse potential.

CORRESPONDING AUTHOR:
Kasra Hosseini
The Alan Turing Institute, London, UK
khosseini@turing.ac.uk

KEYWORDS:
language model; BERT; word2vec; fastText; nineteenth-century English; digital heritage

TO CITE THIS ARTICLE:
Hosseini, K., Beelen, K., Colavizza, G., & Coll Ardanuy, M. (2021). Neural Language Models for Nineteenth-Century English. *Journal of Open Humanities Data*, 7: 22, pp. 1–6. DOI: *https://doi.org/10.5334/johd.48*

1 OVERVIEW

As language is subject to continuous change, the computational analysis of digital heritage should attune models and methods to the specific historical contexts in which these texts emerged. This paper aims to facilitate the "historicization" of Natural Language Processing (NLP) methods by releasing various language models trained on a 19th-century book collection. These models can support research in digital and computational humanities, history, computational linguistics and the cultural heritage or GLAM sector (galleries, libraries, archives, and museums). To accommodate different research needs, we release a wide variety of models, from word type embeddings (word2vec and fastText) to more recent language models that produce context-dependent word or string embeddings (BERT and Flair, respectively). Word type embeddings generate a single vector for a token, regardless of the textual context in which the token appears. On the other hand, "contextual" models generate a distinct token embedding according to the textual context at inference time.

Repository location The dataset is available on Zenodo at *http://doi.org/10.5281/zenodo.4782245*.

Context This work was produced as part of Living with Machines (LwM),[1] an interdisciplinary project focused on the lived experience of Britain's industrialization during the long 19th century. The language models presented here have been used in several research projects, to assess the impact of optical character recognition (OCR) on NLP tasks (van Strien et al., 2020), to detect atypical animacy (Coll Ardanuy et al., 2020), and for targeted sense disambiguation (Beelen et al., 2021).

2 METHOD

2.1 ORIGINAL CORPUS

The original collection consists of ≈48K digitized books in English, made openly available by the British Library in partnership with Microsoft, henceforth *Microsoft British Library* Corpus (MBL). The digitized books are available as JSON files from the British Library web page.[2] ***Figure 1*** gives an overview of the number of books by publication date. The bulk of the material is dated between 1800 and 1900, with the number of documents steeply rising at the end of the 19th century. Since all copyrights are cleared and the data are in the public domain, they have already become a popular resource for (digital) historians and literary scholars.[3] However, one notable issue with this collection (when used for historical research) is the somewhat opaque selection process of books: while the data provides decent coverage over the 19th century, the exact criteria for inclusion remain unclear and future work might profitably consider assessing the characteristics of this collection in more detail (e.g. Pechenick, Danforth, and Dodds (2015)).

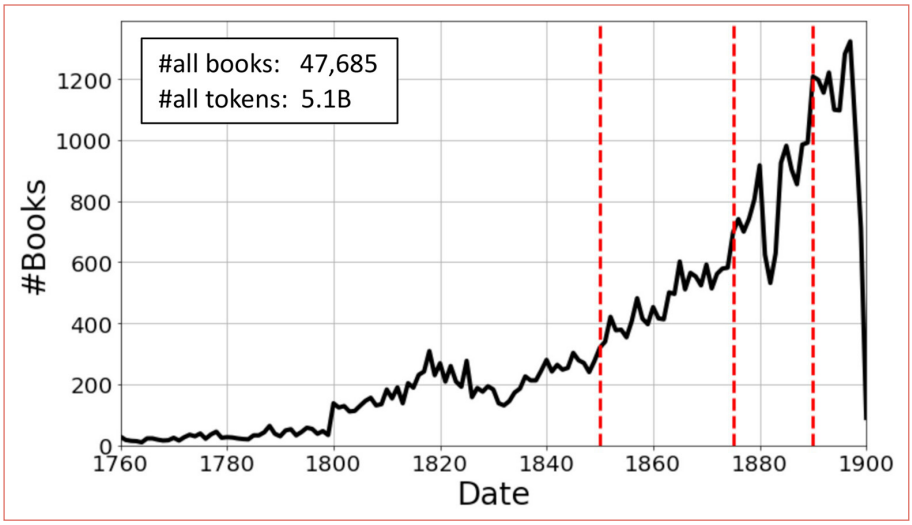

Figure 1 Number of books by publication date. The preprocessed dataset has 47,685 books in English consisting of ≈5.1 billion tokens. The red vertical dashed lines mark the boundaries between the time periods we used to slice the dataset. See Section 2.2 for details.

1 *https://livingwithmachines.ac.uk* (last access: 2021-09-03).

2 *https://data.bl.uk/digbks/db14.html* (last access: 2021-09-03).

3 See, for example, the Contagion Project *https://cca.ucd.ie/contagion-project-british-library-corpus/* (last access: 2021-09-03).

2.2 STEPS

Preprocessing Each book was minimally normalized: we converted the text to ASCII, fixed common punctuation errors, dehyphenated broken tokens, removed most punctuation and separated the remaining punctuation marks from tokens. While the large majority of books in the MBL corpus are written in English, the collection still contains a substantial amount of documents in other languages. Therefore, we filtered by English language, using `spaCy`'s language detector (Honnibal, Montani, Van Landeghem, & Boyd, 2020). Finally, we used `syntok`[4] to split the book into sentences and tokenize the text. This process resulted in one file per book where each line corresponded to a sentence with space-separated tokens.[5]

Data selection For each model architecture, we trained an instance using the whole dataset (i.e., books from all over the 19th century; see ***Figure 1***). For the word2vec and fastText models, we have also trained instances on text published before 1850. Moreover, for BERT, we have fine-tuned four model instances on different time slices, with data from before 1850, between 1850 and 1875, between 1875 and 1890, and between 1890 and 1900, each slice containing ≈1.3B tokens per period, except for 1890–1900, which included ≈1.1B tokens. While this periodization was largely motivated by the number of tokens, the different models (that resulted from the data partitioning) may enable historians to track cultural changes over the long 19th century.[6]

Word2vec and fastText We trained the word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and fastText (Bojanowski, Grave, Joulin, & Mikolov, 2016) models as implemented in the Gensim library (Rehurek & Sojka, 2011). In addition to the preprocessing steps described above, we lowercased all tokens before training. For word2vec, we used the skip-gram architecture, which we trained for one epoch.[7] We set the dimension of the word embedding vectors to 300 and removed tokens appearing less than 20 times. The same hyperparameters were used for training fastText models.[8]

Flair Flair is a character language model based on the Long Short-Term Memory (LSTM) variant of recurrent neural networks (Akbik et al., 2019; Hochreiter & Schmidhuber, 1997). Even though less popular than the Transformers, it has been shown to obtain state-of-the-art results in Named Entity Recognition (NER). We trained a character-level, forward-pass Flair language model on all the books in the MBL corpus for one epoch and sequence length of 250 characters (during training). We used the default character dictionary in Flair. The LSTM component had one layer and a hidden dimension of 2048.[9]

BERT To fine-tune BERT model instances, we started with a contemporary model: 'BERT base uncased',[10] hereinafter referred to as *BERT-base* (Devlin, Chang, Lee, & Toutanova, 2019; Wolf et al., 2019). This instance was then fine-tuned on the earliest time period (i.e., books predating 1850). For the consecutive period (1850–1875), we used the pre-1850 language model instance as a starting point and continued fine-tuning with texts from the following period. This procedure of consecutive incremental fine-tuning was repeated for the other two time periods.

We used the original BERT-base tokenizer as implemented by Hugging Face[11] (Wolf et al., 2019). We did not train new tokenizers for each time period. This way, the resulting language model instances can be compared easily with no further processing or adjustments. The tokenized

4 *https://pypi.org/project/syntok* (last access: 2021-09-03).

5 We evaluated the performance of preprocessing steps (e.g. the tokenizer and language detection components) by manually checking the resulting texts.

6 For example, the pre-1850 dataset sets apart the First Industrial Revolution from later developments in Britain. Likewise, 1890–1900 is set off, especially in literary terms, by the emergence of 'modernist' sensibilities and the questioning of class and gender hierarchies associated with the term '*fin de siècle*'.

7 We followed the procedure proposed by Mikolov et al. (2013) since the data we used for training was comparable in size. Mikolov et al. (2013) started with three epochs and later concluded that one suffices. We will release the scripts used to train our word2vec models, allowing users to select different hyperparameters.

8 The word2vec and fastText models are not directly comparable. Scholars interested in quantifying semantic change between models are encouraged to either compare neighbours (following Gonen, Jawahar, Seddah, and Goldberg (2020)) or align the models, for which there are multiple libraries available. See, for example, *https://github.com/Garrafao/LSCDetection* and *https://github.com/artetxem/vecmap*.

9 Refer to *https://github.com/Living-with-machines/histLM/blob/main/JOHD_paper* for additional information on the neural language models and their preprocessing and training steps.

10 *https://github.com/google-research/bert* (last access: 2021-09-03).

11 *https://github.com/huggingface/transformers* (last access: 2021-09-03).

and lowercased sentences were fed to the language model fine-tuning tool in which only the masked language model (MLM) objective was optimized. We used a batch size of 5 per GPU and fine-tuned for 1 epoch over the books in each time-period. The choice of batch size was dictated by the available GPU memory (we used 4 × NVIDIA Tesla K80 GPUs in parallel). Similar to the original BERT pre-training procedure, we used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and L_2 weight decay of 0.01. In our fine-tuning procedure, we used a linear learning-rate warm-up over the first 2,000 steps. A dropout probability of 0.1 was applied in all layers.

Quality control The quality of our language models was evaluated on multiple downstream tasks. In van Strien et al. (2020), we investigated the impact of OCR quality on the 19th-century word2vec model and showed how language models trained on large OCR'd corpora still yield robust word embedding vectors. However, OCR errors can be unevenly distributed over time and potentially distort the comparison of language models. The BERT models have been used in Coll Ardanuy et al. (2020) and Beelen et al. (2021), where they generally improved the performance of various downstream tasks when the data of the experiments was contemporaneous to that of the language models, thereby confirming their quality via extrinsic evaluation.

Lastly, we need to stress that our language models can contain historical stereotypes and prejudices (related to race, gender, or sexual orientation, among others). We did not attempt to quantify or remove these biases. Therefore, these models should be used critically and responsibly, to avoid the propagation of historical biases (see also Hengchen and Tahmasebi (2021)).

3 LANGUAGE MODEL ZOO

Object name `histLM`.

Format names and versions The models are shared as ZIP files (one per model architecture). The directory structure is described in the `README.md` file.

Creation dates 2020-01-31 to 2020-10-07.

Dataset creators Kasra Hosseini, Kaspar Beelen and Mariona Coll Ardanuy (The Alan Turing Institute) preprocessed the text, created a database, trained and fine-tuned language models as described in this paper. Giovanni Colavizza (University of Amsterdam) initiated this work on historical language models. All authors contributed to planning and designing the experiments.

Language The language models have been trained on 19th-century texts in English.

License The models are released under open license CC BY 4.0, available at *https://creativecommons.org/licenses/by/4.0/legalcode*.

Repository name All the language models are published in Zenodo at *http://doi.org/10.5281/zenodo.4782245*. We have also provided scripts to work with the language models, available on GitHub at *https://github.com/Living-with-machines/histLM*.

Publication date 2021-05-23.

4 REUSE POTENTIAL

Even though word2vec has been around for almost a decade—an eternity in the fast-moving NLP ecosystem—the word type embeddings it produces persist as popular instruments, especially for interdisciplinary research (Azarbonyad et al. 2017; Hengchen, Ros, & Marjanen, 2019). The more recent fastText model extends on word2vec by using subword information. Contextualized language models have meant a breakthrough in NLP research (e.g. Smith (2019) for an overview), as they represent words in the contexts in which they appear, instead of conflating all senses, one of the main criticisms of word type embeddings. The potential of using such models for historical research is immense as they allow a more accurate context-dependent representation of meaning. These embeddings can also be used in existing tools for historical research (e.g. Hosseini, Nanni, and Coll Ardanuy (2020)).

Given that existing libraries, such as Gensim, Flair, or Hugging Face, provide convenient interfaces to work with these embeddings, we are confident that our historical models will serve the needs of a wide-variety of scholars, from NLP and data science to the humanities,

for different tasks and research purposes, such as measuring how words change meaning over time (Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015; Tahmasebi, Borin, & Jatowt, 2018), automatic OCR correction (Hämäläinen & Hengchen, 2019), interactive query expansion[12] or, more generally, any research that involves diachronic language change.

ACKNOWLEDGEMENTS

We thank Nilo Pedrazzini and three anonymous reviewers for their careful and constructive reviews. We are grateful to David Beavan (The Alan Turing Institute) and James Hetherington (University College London) for helping with the data access and research infrastructure. We thank the British Library for supplying digitised books.

FUNDING STATEMENT

This work was supported by Living with Machines (AHRC grant AH/S01179X/1) and The Alan Turing Institute (EPSRC grant EP/N510129/1). Living with Machines, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with The Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Kasra Hosseini *orcid.org/0000-0003-4396-6019*
The Alan Turing Institute, London, UK

Kaspar Beelen *orcid.org/0000-0001-7331-1174*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

Giovanni Colavizza *orcid.org/0000-0002-9806-084X*
Institute for Logic, Language and Computation, University of Amsterdam, NL

Mariona Coll Ardanuy *orcid.org/0000-0001-8455-7196*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

REFERENCES

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & **Vollgraf, R.** (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 annual conference of the north american chapter of the association for computational linguistics (demonstrations)* (pp. 54–59).

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & **Kamps, J.** (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 1509–1518). DOI: *https://doi.org/10.1145/3132847.3132878*

Beelen, K., Nanni, F., Coll Ardanuy, M., Hosseini, K., Tolfo, G., & **McGillivray, B.** (2021). When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of acl-ijcnlp*. Bangkok, Thailand (Online): Association for Computational Linguistics. DOI: *https://doi.org/10.18653/v1/2021.findings-acl.243*

Bojanowski, P., Grave, E., Joulin, A., & **Mikolov, T.** (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*. DOI: *https://doi.org/10.1162/tacl_a_00051*

Coll Ardanuy, M., Nanni, F., Beelen, K., Hosseini, K., Ahnert, R., Lawrence, J., ..., & **McGillivray, B.** (2020, December). Living machines: A study of atypical animacy. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4534–4545). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2020.coling-main.400.* DOI: *https://doi.org/10.18653/v1/2020.coling-main.400*

12 See, for example, the search tools provided by the Impresso interface *https://impresso-project.ch* (last access: 2021-09-03).

Devlin, J., Chang, M.-W., Lee, K., & **Toutanova, K.** (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/N19-1423.* DOI: *https://doi.org/10.18653/v1/N19-1423*

Gonen, H., Jawahar, G., Seddah, D., & **Goldberg, Y.** (2020). 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* (pp. 538–555). DOI: *https://doi.org/10.18653/v1/2020.acl-main.51*

Hämäläinen, M., & **Hengchen, S.** (2019). From the paft to the fiiture: a fully automatic nmt and word embeddings method for ocr post-correction. *arXiv preprint arXiv:1910.05535*. DOI: *https://doi.org/10.26615/978-954-452-056-4_051*

Hengchen, S., Ros, R., & **Marjanen, J.** (2019). A data-driven approach to the changing vocabulary of the nation in english, dutch, swedish and finnish newspapers, 1750–1950. In *Proceedings of the digital humanities (dh) conference*.

Hengchen, S., & **Tahmasebi, N.** (2021). A collection of swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data*, *7*. DOI: *https://doi.org/10.5334/johd.22*

Hochreiter, S., & **Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. DOI: *https://doi.org/10.1162/neco.1997.9.8.1735*

Honnibal, M., Montani, I., Van Landeghem, S., & **Boyd, A.** (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. DOI: *https://doi.org/10.5281/zenodo.1212303*

Hosseini, K., Nanni, F., & **Coll Ardanuy, M.** (2020, October). DeezyMatch: A flexible deep learning approach to fuzzy string matching. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 62–69). Online: Association for Computational Linguistics. Retrieved from *https://aclanthology.org/2020.emnlp-demos.9.* DOI: *https://doi.org/10.18653/v1/2020.emnlp-demos.9*

Kingma, D. P., & **Ba, J.** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & **Skiena, S.** (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web* (pp. 625–635). DOI: *https://doi.org/10.1145/2736277.2741627*

Mikolov, T., Chen, K., Corrado, G., & **Dean, J.** (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*. Retrieved 2019-11-20, from *http://arxiv.org/abs/1301.3781*

Pechenick, E. A., Danforth, C. M., & **Dodds, P. S.** (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, *10*(10), e0137041. DOI: *https://doi.org/10.1371/journal.pone.0137041*

Rehurek, R., & **Sojka, P.** (2011). Gensim-python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Smith, N. A. (2019). Contextual word representations: A contextual introduction. *arXiv preprint arXiv:1902.06006*.

Tahmasebi, N., Borin, L., & **Jatowt, A.** (2018). Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., & **Colavizza, G.** (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *Icaart*, *1*, 484–496. DOI: *https://doi.org/10.5220/0009169004840496*

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ..., & **Brew, J.** (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv, abs/1910.03771*. DOI: *https://doi.org/10.18653/v1/2020.emnlp-demos.6*

TO CITE THIS ARTICLE:
Hosseini, K., Beelen, K., Colavizza, G., & Coll Ardanuy, M. (2021). Neural Language Models for Nineteenth-Century English. *Journal of Open Humanities Data*, 7: 22, pp. 1–6. DOI: *https://doi.org/10.5334/johd.48*

Published: 27 September 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.