A Collection of Swedish Diachronic Word Embedding Models Trained on Historical Newspaper Data

SIMON HENGCHEN • NINA TAHMASEBI

*Author affiliations can be found in the back matter of this article

ABSTRACT

This paper describes the creation of several word embedding models based on a large collection of diachronic Swedish newspaper material available through Språkbanken Text, the Swedish language bank. This data was produced in the context of Språkbanken Text’s continued mission to collaborate with humanities and natural language processing (NLP) researchers and to provide freely available language resources, for the development of state-of-the-art NLP methods and tools.

CORRESPONDING AUTHOR:
Simon Hengchen
Språkbanken Text, Department of Swedish, University of Gothenburg, SE
simon.hengchen@gu.se

KEYWORDS:
word embeddings; semantic change; newspapers; diachronic word embeddings

TO CITE THIS ARTICLE:
1 OVERVIEW

We release diachronic word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) models in their skip-gram with negative sampling (SGNS) architecture. The models are trained on 20-year time bins, with two temporal alignment strategies: independently-trained models for post-hoc alignment (as introduced by Kulkarni et al. 2015), and incremental training (Kim et al., 2014). In the incremental scenario, a model for $t_1$ is trained and saved, then updated with the data from $t_2$. The resulting model is in turn saved, then updated with data from $t_3$, etc. Given that space alignment is shown to be noisy (Dubossarsky et al., 2017, 2019), we release the independently trained models without alignment and leave the choice of alignment algorithm to the end user.  

To make this data release useful out-of-the-box, as well as to foster reuse by researchers from various fields, we release models, code to run the whole pipeline, code examples to load and use the models, and documentation. 

We would like to underline that language change is driven by humans. Humans learn from their mistakes and what was once considered acceptable thankfully is not anymore. Machine learning models trained on data from the past inevitably learn biases of their time and as a result, the models shared with this paper contain characteristics of the past. These characteristics include sexism, racism, antisemitism, homophobia, and other types of unacceptable characteristics of their time (see e.g. Tripodi et al., 2019). The authors do not endorse them, and neither does the University of Gothenburg. Nonetheless, whilst this is not the aim of this paper, we hope it can also help shed light on these representations, as ignoring them would mean they have never existed.

1.1 REPOSITORY LOCATION

The data set is available on Zenodo at https://zenodo.org/record/4301658.

1.2 CONTEXT

This data was produced in the context of Språkbanken Text’s continued mission to collaborate with humanities and natural language processing (NLP) researchers and to provide freely available language resources for the development of state-of-the-art NLP methods and tools.

2 METHOD

We retrieved all Sparv-pipeline (Borin et al., 2016) processed XML files in the Kubhist 2 newspaper archive: Språkbanken Text makes several corpora – including Kubhist 2 – available through web interface Korp (Borin et al., 2012). The data in Korp has been processed (dependency parsing, semantic annotation, lemmatisation, etc.) with the Sparv pipeline. The original dataset and the specific steps are described below.

2.1 ORIGINAL DATA

The entirety of the Kungliga bibliotekets historiska tidningar (‘The Royal Library’s historical newspapers,’ Kubhist 2) corpus (Språkbanken, 2019) was used. For a detailed description of the corpus, we refer to Adesam et al. (2019) and to a blog post by Dana Dannêl. ‘Kubhist 2 contains over 5.5 billion tokens, and it is made up of newspapers from all over Sweden.

2.2 STEPS

- Extracted all words from the XML
- Given the relative quality of the optical character recognition (OCR) output and to reduce the amount of OCR errors in the data set, we cleaned the resulting text with the following procedure:

1 We refer to the LSCDetection repository (Schlechtweg et al., 2019; Schlechtweg & Schulte im Walde, 2020) for a selection of alignment methods: https://github.com/Garrafao/LSCDetection/tree/master/alignment.


3 For a discussion on how OCR can affect quantitative text analyses, we refer the reader to Hill & Hengchen (2019) and van Strien et al. (2020).
- lowercasing;
- removing digits;
- removing all characters not belonging to the (lowercased) Swedish alphabet, which consists of the 26 letters in the Latin alphabet and å, ä, ö. This includes the removal of punctuation marks;
- removing tokens the length of which is two characters or smaller

• Joined files belonging to the same double decade, starting with our earliest time bin of 1740 and ending in 1899 (i.e. 1740–1750; 1760–1770; ...; 1880–1890, where e.g. 1740 covers 1740; 1741; ...; 1749)

• For each time bin, trained two type-embedding language models with two “alignment” strategies:
  - word2vec, independently-trained and incrementally trained
  - fastText, independently-trained and incrementally trained

For both language models, we use the default hyperparameters in gensim (Réhurek & Sojka, 2010) aside from: vector dimensionality of 100, frequency threshold of 50, seed of 1830. The choice for default parameters is explained in Subsection 2.3.

2.3 QUALITY CONTROL

Several sanity checks were made during the preprocessing of the original data, including:

• Manual matching of records between the original XML and resulting corpus files as well as against the Korp version of Kubhist 2;

• Manual matching of metadata between the original XML and resulting corpus files as well as against the Korp version of Kubhist 2.

It is notoriously difficult to evaluate the quality of word embeddings for historical data, as annotated test sets are either lacking or extremely costly (Schlechtweg et al., 2020; Hengchen et al., 2021b). While synthetic evaluation procedures exist (Cook & Stevenson, 2010; Kulkarni et al., 2015; Rosenfeld & Erk, 2018; Dubossarsky et al., 2019; Shoemark et al., 2019), they are tailored for the specific task of semantic change (usually, the task of determining if there is a change of a word’s meaning over time) and are not suited for general-purpose diachronic word embeddings as they might lead to privileging a (set of) hyperparameter(s) that might be detrimental to other tasks. As a result, we use default parameters and carry out a small-scale quality control by a) verifying that the code written does what it is expected to do; and b) making sure that models output semantic similarity, as expected.

The code to train word embeddings was read (but not run) by two computational linguists who have extensive experience with diachronic word embeddings and are not authors of this paper, and no errors were found. Once the word embeddings were trained, we selected several target words and systematically extracted the most similar terms for every model trained. Similar terms were then evaluated by a native speaker of Swedish who confirmed that such terms were indeed, to the best of their knowledge, semantically similar. In many cases and especially so for the fastText models that harvest subword information, the most similar words consist of OCR errors and spelling variations, an interesting avenue to pursue in future research. A (non-native speaker of Swedish) reviewer, whom we thank, also performed checks on the local neighbourhoods of selected terms as well as vector arithmetics, and confirmed the models behaved as expected.

We would like to note that the first time bins are very scarce in data, and warn researchers that previous work indicates this has a large influence on the stability of nearest-neighbour distances (Antoniak & Mimno, 2018). We would also like to acknowledge that different temporal alignment strategies might benefit from different hyperparameters for specific tasks (see e.g. I.e.: $\alpha = 0.025$, $\text{window} = 5$, sample = 0.001, min_alpha = 0.0001, negative = 5.
Kaiser et al. 2020 for vector dimensionality in lexical semantic change (LSC)), however, for the reasons stated above we do not perform any tuning on any specific task either.

3 DATASET DESCRIPTION

3.1 OBJECT NAME

The dataset is named HENGCHEN-TAHMASEBI_-_2020_-_Kubhist2_diachronic_embeddings.zip.

3.2 FORMAT NAMES AND VERSIONS

The data is shared as a ZIP file containing gensim binary files (.ft for fastText models, .w2v for word2vec models) and Python (.py) scripts. For the larger models, matrices and vectors are stored separately as NumPy arrays (Oliphant, 2006, .npy). Given the relatively large size of the archive, we recommend that Windows users decompress the file (right-click → ‘Extract all’) instead of double-clicking it. The directory structure is as follows:

```
ROOT/
   README.md
   code/
      *.py files
      requirements.txt
   fasttext/
      incremental/
         *.ft files
         *.npy files
      indep/
         *.ft files
         *.npy files
   word2vec/
      incremental/
         *.w2v files
         *.npy files
      indep/
         *.w2v files
         *.npy files
```

The README.md file contains basic information about this release, while the code/requirements.txt file contains a list of required Python packages to run the provided code.

3.3 CREATION DATES

The models were trained on 2020-09-15.

3.4 DATASET CREATORS

The original data was scanned and OCRed by the National Library of Sweden. It consists of Swedish newspapers from all parts of Sweden. It has since been run through the Sparv annotation pipeline by Martin Hammarstedt at Språkbanken Text. As described in Subsection 2.2 the authors of this paper have extracted the text from the original XML, processed it, and trained the models.

3.5 LANGUAGE

The diachronic word embedding models have been trained on Swedish data. The variable names in the accompanying Python code and documentation are in English.
3.6 LICENSE
The models and code are released under open license CC BY 4.0, available at https://creativecommons.org/licenses/by/4.0/legalcode.

3.7 REPOSITORY NAME
The data is released on Zenodo, and named ‘A collection of Swedish diachronic word embedding models trained on historical newspaper data.’ A link to the Zenodo repository as well as a description of the dataset are also available on the Språkbanken Text website, along with other resources.

3.8 PUBLICATION DATE
The data was released on Zenodo on 2020/12/2.

4 REUSE POTENTIAL
We believe that this data release can be re-used by a relatively large community of researchers from different fields. This fact is reinforced by the release of documented code – bypassing the need for advanced technical skills, which is one of the key challenges in interdisciplinary collaborations (McGillivray et al., 2020).

Since the models span several decades, they present an interesting view of words over time, useful for researchers interested in diachronic studies such as culturomics (Michel et al., 2011), semantic change (see Tahmasebi et al. (2018); Kutuzov et al. (2018), for overviews), historical research (van Eijnatten & Ros, 2019; Hengchen et al., 2021a; Marjanen et al., 2020), etc. They also can be further fed as input to more complex neural networks tackling downstream tasks aimed at historical data such as OCR post-correction (Hämäläinen & Hengchen, 2019; Duong et al., 2020) or more linguistics-oriented problems (Budts, 2020). Since we release the whole models and not solely the learned vectors, these models can be further trained and specialised, or used by NLP researchers to compare different space alignment procedures.

ACKNOWLEDGEMENTS
The authors would like to thank the researchers and developers at Språkbanken Text for creating and releasing the processed XML for Kubhist 2 used as a basis for training the embeddings. Additional thanks go to Haim Dubossarsky and Dominik Schlechtweg for their insightful comments on the robustness of the pipeline that trains word embeddings. Finally, our sincere thanks go to both reviewers for their useful remarks.

FUNDING STATEMENT
This work has been funded in part by the project Towards Computational Lexical Semantic Change Detection supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and Nationella Språkbanken (the Swedish National Language Bank) – jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and its 10 partner institutions, to Nina Tahmasebi. The authors wish to thank the Department of Swedish at the University of Gothenburg for providing financial support for the Open Access fee.

COMPETING INTERESTS
The authors have no competing interests to declare.

AUTHOR AFFILIATIONS
Simon Hengchen orcid.org/0000-0002-8453-7221
Språkbanken Text, Department of Swedish, University of Gothenburg, SE

Nina Tahmasebi orcid.org/0000-0003-1688-1845
Språkbanken Text, Department of Swedish, University of Gothenburg, SE
REFERENCES


TO CITE THIS ARTICLE:

Published: 27 January 2021

COPYRIGHT:
© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

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