



# Empirical Study of Diachronic Word Embeddings for Scarce Data

Syrielle Montariol, Alexandre Allauzen

## ► To cite this version:

Syrielle Montariol, Alexandre Allauzen. Empirical Study of Diachronic Word Embeddings for Scarce Data. Recent Advances in Natural Language Processing, Sep 2019, Varna, Bulgaria. pp.795-803, 10.26615/978-954-452-056-4\_092 . hal-04421484

**HAL Id: hal-04421484**

**<https://hal.science/hal-04421484v1>**

Submitted on 27 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

# Empirical Study of Diachronic Word Embeddings for Scarce Data

Syrielle Montariol

LIMSI - CNRS

Univ. Paris-Sud, Univ. Paris-Saclay

Société Générale

syrielle.montariol@limsi.fr

Alexandre Allauzen

LIMSI - CNRS

Univ. Paris-Sud, Univ. Paris-Saclay

alexandre.allauzen@limsi.fr

## Abstract

Word meaning change can be inferred from drifts of time-varying word embeddings. However, temporal data may be too sparse to build robust word embeddings and to discriminate significant drifts from noise. In this paper, we compare three models to learn diachronic word embeddings on scarce data: incremental updating of a Skip-Gram from Kim et al. (2014), dynamic filtering from Bamler and Mandt (2017), and dynamic Bernoulli embeddings from Rudolph and Blei (2018). In particular, we study the performance of different initialisation schemes and emphasise what characteristics of each model are more suitable to data scarcity, relying on the distribution of detected drifts. Finally, we regularise the loss of these models to better adapt to scarce data.

## 1 Introduction

In all languages, word usage can evolve over time, mirroring cultural or technological evolution of society (Aitchison, 2001).

For example, the word "Katrina" used to be exclusively a first name until year 2005 when hurricane Katrina devastated the United States coasts. After that tragedy, this word started to be associated with the vocabulary of natural disasters.

In linguistics, *diachrony* refers to the study of temporal variations in the use and meaning of a word. Detecting and understanding these changes can be useful for linguistic research, but also for many tasks of Natural Language Processing (NLP). Nowadays, a growing number of historical textual data is digitised and made publicly available. It can be analysed in parallel with contemporary documents, for tasks ranging from text classification to information retrieval. However, the use of conventional word embeddings methods have the drawback to average in one vector the different word's usages observed across the whole corpus.

This *static* representation hypothesis turns out to be limited in the case of temporal datasets.

Assuming that a change in the context of a word mirrors a change in its meaning or usage, a solution is to explore diachronic word embeddings: word vectors varying through time, following the changes in the global context of the word. While many authors proposed diachronic embedding models these last years, these methods usually need large amounts of data to ensure robustness.

However, temporal datasets often face the problem of scarcity; beyond the usual scarcity problem of domain-specific corpora or low-resource languages, a temporal dataset can have too short compared to the volume of the full dataset.<sup>1</sup> Moreover the amount of digital historical texts is limited for many languages, particularly for oldest time periods.

This paper addresses the following question: In case of scarce data, how to efficiently learn time-varying word embeddings? For this purpose, we compare three diachronic methods on several sizes of datasets. The first method is incremental updating (Kim et al., 2014), where word vectors of one time step are initialised using the vectors of the previous time step. The second one is the dynamic filtering algorithm (Bamler and Mandt, 2017) where the evolution of the embeddings from one time step to another is controlled using a Gaussian diffusion process. Finally, we experiment dynamic Bernoulli embeddings (Rudolph and Blei, 2018) where the vectors are jointly trained on all time slices.

These three models are briefly described in section 3. The hyper-parameters are specifically tuned towards efficiency on small datasets. Then, we explore the impact of different initialisation

<sup>1</sup>A short period can be one month or less, depending on the domain.

scheme and compare the behaviour of word drifts exhibited by the models. Finally, we experiment regularising the models in order to tackle the faults detected in the previous analysis. The experiments in section 4 are made on the *New York Times Annotated Corpus* (NYT)<sup>2</sup> (Sandhaus, 2008) covering two decades.

## 2 Related Work

The first methods to measure semantic evolution rely on detecting changes in word co-occurrences, and approaches based on distributional similarity (Gulordava and Baroni, 2011). The use of automated learning methods, based on word embeddings (Mikolov et al., 2013), is recent and has undergone a increase in interest these last two years with the successive publication of three articles dedicated to a literature review of the domain (Kutuzov et al., 2018; Tahmasebi et al., 2018; Tang, 2018). In this section, we mainly consider this second line of work, along with the peculiarities of scarce data.

Kim et al. (2014) developed one of the first method to learn time-varying word sparse representations. It consists in learning an embedding matrix for the first time slice  $t_0$ , then updating it at each time step  $t$  using the matrix at  $t - 1$  as initialisation. This method is called incremental updating. Another broadly used method is to learn an embedding matrix for each time slice independently; due to the stochastic aspect of word embeddings, the vectorial space for each time slice is different, making them not directly comparable. Thus, authors perform an alignment of the embeddings spaces by optimising a geometric transformation (Hamilton et al., 2016; Dubossarsky et al., 2017; Szymanski, 2017; Kulkarni et al., 2015)).

In the case of sparse data, in addition to the approximative aspect of the alignment that harms the robustness of the embeddings, these methods are sensitive to random noise, which is difficult to disambiguate from semantic drifts. Moreover, the second one requires large amounts of data for each time step to prevent overfitting. Tahmasebi (2018) shows that low-frequency words have a much lower temporal stability than high-frequency ones. In (Tahmasebi et al., 2018), the authors explain that usual methods for diachronic embeddings

training such as the two previously presented are inefficient in the case of low-frequency words and hypothesise that a new set of methods, pooled under the name of *dynamic* models, may be more adapted. These models use probabilistic models to learn time-varying word embeddings while controlling the drift of the word vectors using a Gaussian diffusion process. Bamler and Mandt (2017) uses Bayesian word embeddings, which makes the algorithm more robust to noise in the case of sparse data; while Rudolph and Blei (2018) relies on a Bernoulli distribution to learn the dynamic embeddings jointly across all time slices, making the most of the full dataset.

Outside of the framework of diachrony, several attempts aim at improving or adapting word embeddings to low-volume corpora in the literature. It can involve morphological information (Luong et al., 2013) derived from the character level (Santos and Zadrozny, 2014; Labeau et al., 2015), and often make use of external resources: semantic lexicon (Faruqui et al., 2015), and pre-trained embeddings from larger corpora (Komiya and Shinou, 2018). However, to our knowledge, no work has attempted to apply similar solutions to the problem of sparse data in temporal corpora, even though this situation has been faced by many authors, often in the case of short time steps for social media data (Stewart et al., 2017; Bamler and Mandt, 2017; Kulkarni et al., 2015).

## 3 Diachronic Models

This section briefly describes the three models under study: the Skip-Gram incremental updating algorithm from Kim et al. (2014), the dynamic filtering algorithm of Bamler and Mandt (2017), and the dynamic Bernoulli embeddings model from Rudolph and Blei (2018). We consider a corpus divided into  $T$  time slices indexed by  $t$ . For each time step  $t$ , every word  $i$  is associated with two vectors  $u_{it}$  (word vector) and  $v_{it}$  (context vector).

### 3.1 Incremental Skip-Gram (ISG)

This algorithm relies on the skip-gram model estimated with negative sampling (SGNS) method described in (Mikolov et al., 2013) and it can be summarised as follows. The probability of a word  $i$  to appear in the context of a word  $j$  is defined by  $\sigma(u_{i,t}^T v_{j,t})$ , with  $\sigma$  being the sigmoid function. Words  $i$  and  $j$  are represented by their embedding

<sup>2</sup><https://catalog.ldc.upenn.edu/LDC2008T19>

vectors  $u_{i,t}$  and  $v_{j,t}$ ) at time  $t$ . The matrices  $U_t$  and  $V_t$  gathers all of them for the whole vocabulary. The context is made of a fixed number of surrounding words and each word in the context are considered as independent of each other.

The negative sampling strategy associates to each observed word-context pair (the positive examples)  $n_{ijt}^+$ , a set of negative examples  $n_{ijt}^-$ . The negative examples are sampled for a noise distribution following Mikolov et al. (2013).

Let  $n_t^{+-}$  denote for the time step  $t$ , the union of positive and negative examples. The objective function can be defined as the following log-likelihood:

$$\begin{aligned} \log p(n_t^{+-}|U_t, V_t) &= \mathcal{L}_{pos}(U_t, V_t) + \mathcal{L}_{neg}(U_t, V_t) \\ &= \sum_{i,j=1}^L (n_{ijt}^+ \log \sigma(u_{i,t}^T v_{j,t}) + n_{ijt}^- \log \sigma(-u_{i,t}^T v_{j,t})) \end{aligned} \quad (1)$$

For the first time slice, the matrices  $U_1$  and  $V_1$  are initialised using a Gaussian random noise  $\mathcal{N}(0, 1)$  before being trained according to equation (1). Then, for each successive time slice, the embeddings are initialised with values of the previous time slice following the methodology of (Kim et al., 2014). This way, the word vectors of each time step are all in the same vectorial space and directly comparable.

### 3.2 Dynamic Filtering of Skip-Gram (DSG)

This second method relies on the Bayesian extension of the SGNS model described by Barkan (2015). The main idea is to share information from one time step to another, allowing the embeddings to drift under the control of a diffusion process. A full description of this approach, denoted as the filtering model, can be found in (Bamler and Mandt, 2017).

In this model, the vectors  $u_{i,t}$  and  $v_{i,t}$  are considered as latent vectors. Under a Gaussian assumption, they are represented by their means  $(\mu_{u_{i,t}}, \mu_{v_{i,t}})$  and variances  $(\Sigma_{u_{i,t}}, \Sigma_{v_{i,t}})$ . They are initialised for the first time slice with respectively a zero mean vector and a identity variance matrix.

The temporal drift from one time step to another follows a Gaussian diffusion process with zero mean and variance  $D$ . This variance is called the *diffusion* constant and has to be tuned along with the other hyperparameters. Moreover, at each time step a second Gaussian prior with zero mean

and variance  $D_0$  is added, resulting in the following distributions over the embeddings matrices  $U_t$ :

$$\begin{aligned} p(U_1|U_0) &\sim \mathcal{N}(0, D_0) \\ p(U_t|U_{t-1}) &\sim \mathcal{N}(U_{t-1}, D) \mathcal{N}(0, D_0). \end{aligned} \quad (2)$$

The same equation stands for  $V_t$ . Training this model requires to estimate the posterior distributions over  $U_t$  and  $V_t$  given  $n_t^{+-}$ . This (Bayesian) inference step is unfortunately untractable. In (Bamler and Mandt, 2017), the authors propose to use variational inference (Jordan et al., 1999) in its online extension (Blei et al., 2017). The principle of variational inference is to approximate the posterior distribution with a simpler variational distribution  $q_\lambda(U, V)$  ( $\lambda$  gathers all the parameters of  $q$ ). This variational posterior will be iteratively updated at each time step. The final objective function can be written as follows:

$$\begin{aligned} \mathcal{L}_t(\lambda) &= E_{q_\lambda}[\log p(n_t^{+-}|U_t, V_t)] \\ &\quad + E_{q_\lambda}[\log p(U_t, V_t)|n_{1:t-1}^{+-}] \\ &\quad - E_{q_\lambda}[\log q_\lambda(U_t, V_t)]. \end{aligned} \quad (3)$$

This loss function is the sum of three terms: the log-likelihood (computed following equation (1)), the log-prior (which enforces the smooth drift of embedding vectors, sharing information with the previous time step), and the entropy term (approximated as the sum of the variances of the embedding vectors).

### 3.3 Dynamic Bernoulli Embeddings (DBE)

The DBE models extends the *Exponential Family Embeddings* (EFE)(Rudolph et al., 2016), a probabilistic generalisation of the *Continuous Bag-of-Words* (CBOW) model of Mikolov et al. (2013). The main idea is that the model predicts the central word vector conditionally to its context vector following a Bernoulli distribution. A detailed description of the model can be found in (Rudolph and Blei, 2018).

Each word  $i$  has  $T$  different embeddings vectors  $u_{it}$ , but this time, the context vectors  $v_i$  are assumed to be fixed across the whole corpus. The embedding vector  $u_{it}$  drifts throughout time following a Gaussian random walk, very similarly to equation (2):

$$\begin{aligned} U_0, V &\sim \mathcal{N}(0, \lambda_0^{-1}I), \\ U_t &\sim \mathcal{N}(U_{t-1}, \lambda^{-1}I). \end{aligned} \quad (4)$$

The *drift* hyper-parameter  $\lambda$  controls the temporal evolution of  $U_t$ , and is shared across all time steps. The training process, described more precisely by [Rudolph and Blei \(2018\)](#), relies on a variant of the negative sampling strategy described by [Mikolov et al. \(2013\)](#). The goal is to optimise the model across all time steps jointly, by summing over  $t$  the following loss function:

$$\mathcal{L}_t = \mathcal{L}_{pos}(U_t, V) + \mathcal{L}_{neg}(U_t, V) + \mathcal{L}_{prior}(U_t, V). \quad (5)$$

The two first terms are computed as in equation (1). The third term is defined as :

$$\begin{aligned} \mathcal{L}_{prior}(U_t, V) = & -\frac{\lambda_0}{2} \sum_{i=1}^L \|v_i\|^2 - \frac{\lambda_0}{2} \sum_{i=1}^L \|u_{i,0}\|^2 \\ & - \frac{\lambda}{2} \sum_{i,t} \|u_{i,t} - u_{i,t-1}\|^2. \end{aligned} \quad (6)$$

The role of  $\mathcal{L}_{prior}$  is twofold: it acts as a regularisation term on  $V$  and  $U_t$ , and as a constraint on the drift of  $U_t$ , preventing it from going too far apart from  $U_{t-1}$ .

## 4 Experimental Results

The goal of this study is to compare the behaviour of the three algorithms described in section 3 in case of low-volume corpora. We evaluate their predictive power on different volumes of data to compare the impact of two initialisation methods, and analyse the behaviour of the drift of the embeddings.

### 4.1 Experimental Setup

We use the *New York Times Annotated Corpus* (NYT) ([Sandhaus, 2008](#))<sup>3</sup> containing around 1 855 000 articles ranging from January 1<sup>st</sup> 1987 to June 19th 2007. We divide the corpus into  $T = 20$  yearly time steps (the incomplete last year is not used in the analysis) and held out 10 % of each time step for validation and testing. Then, we sample several subsets of the corpus : 50 %, 10%, 5% and 1% of the training set. This way, we can compare the models on each subset to evaluate their ability to train a model in the case of low-volume corpora.

We remove stopwords and choose a vocabulary of  $V = 10k$  most frequent words. Indeed,

a small vocabulary is more adequate for sparse data in a temporal analysis in order to avoid having time steps where some word does not appear at all. The total number of words in the corpus after preprocessing is around 38.5 million. It amounts to around 200k words per time step in the 10 % subset of the corpus, thus only 20k in the 1 % subset.

To tune the hyperparameters, we use the log-likelihood of positive examples  $\mathcal{L}_{pos}$  measured on the validation set. We train each model for 100 epochs, with a learning rate of 0.1, using the Adam optimiser. For the DSG model, we use a diffusion constant  $D = 1$  and a prior variance  $D_0 = 0.1$  for both corpora. For the DBE model, we use  $\lambda = 1$  and  $\lambda_0 = 0.01$ .

We choose an embedding dimension  $d = 100$ , as the experiments show that a small embedding dimension, as in ([Stewart et al., 2017](#)), leads to smoother word drifts and makes the model less sensitive to noise when the data is scarce.

We use a context window of 4 words and a negative ratio of 1; we observed that having a higher number of negative samples artificially increased the held-out likelihood, but equalised the drifts of all the words in the corpus. Thus, in an extreme scarcity situation, each negative sample has a high weight during training: the number of negative samples has to be very carefully selected depending on the amount of data.

### 4.2 Impact of Initialisation on Sparse Data

The embedding vectors of the ISG and DBE models are initialised using a Gaussian white noise, while the means and variances of DSG are initialised with null vectors and identity matrices respectively. However, a good initialisation can greatly improve the quality of embeddings, particularly in the case of scarce data.

We experiment the impact of two types of initialisation on the log-likelihood of positive examples on the test set.

#### Internal initialisation:

We train each model in a static way on the full dataset. Then, we use the resulting vectors as initialisation for the first time step of the diachronic models. This method is especially suitable for domain-specific corpora where no external comparable data is available.

<sup>3</sup>released by the Linguistic Data Consortium



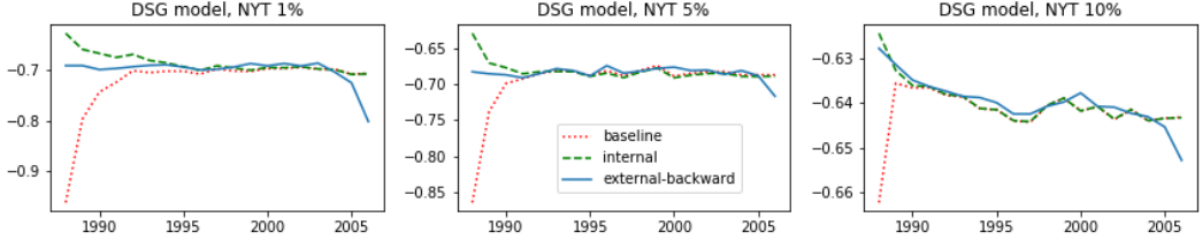


Figure 1: Log-likelihoods for the DSG model on three subsets of the corpus, comparing the baseline (random initialisation) with the two initialisation methods: *internal* is the initialisation from the full dataset while *external-backward* is the initialisation with the Wikipedia vectors, with training from most recent to oldest time step.

### Backward external initialisation:

We use a set of embeddings pre-trained on a much larger corpus for initialisation : The Wikipedia corpus (dump of August 2013) (Li et al., 2017) with vectors of size 100. These embeddings are representative of the use of words in 2013; and in general, large corpora exist almost exclusively for recent periods. Thus, we choose to use the pre-trained embeddings as initialisation for the *last* time step (the most recent). Then, we update the embeddings incrementally from new to old (*reverse incremental updating*).

This method would be particularly suitable for corpora with low volume in older time slices, as it is the case for most of the historical dataset in languages other than English.

For the DSG model, the pre-trained vectors are used as the mean parameter for each word. The variance parameter is fixed at 0.1. Experiments with a prior variance of 0.01 and 1 had a lowest log-likelihood on the validation set.

The log-likelihood curves in figure 1 show that the internal initialisation has a better impact on the likelihood at the beginning of the period, as it is closer to the data than the external initialisation. The positive impact of the backward external initialisation increases with the volume of data.

Overall, the mean log-likelihoods across all time steps (Table 1) are higher using the internal initialisation. We conjecture that internal initialisation is more profitable to the model when the period is short (here, two decades) with low variance. The backward external initialisation has very close scores to the internal one, and is more suitable for higher volume datasets with a longer period, as it gives higher benefit to the likelihood for bigger subsets.

Initialisation / Model	Random	Internal	Backward external
ISG	-3.17	-2.589	-2.686
DSG	-0.749	-0.686	-0.695
DBE	-2.935	-2.236	-2.459

Table 1: Log-likelihood on the 5% subset of the NYT corpus for each model, with the three initialisation schemes.

### 4.3 Visualising Word Drifts

A high log-likelihood performance does not necessarily imply that the drifts detected by the models are meaningful. In this section, we examine the distribution of word drifts outputted by each model with the internal initialisation. The computed drift is the L2-norm of the difference between the embeddings at  $t_0$  and the embeddings at each  $t$ :

$$drift(U_t) = \left[ \sum_{i=1}^L (u_{i,t} - u_{i,t_0})^2 \right]^{1/2} \quad (7)$$

In the case of the DSG model where the words are represented as distributions, we compute the difference of the mean vectors.

We plot the superimposed histograms of *successive* drifts (Figure 2) from  $t_0 = 1987$  to each successive time step, for all studied models. For example, on the histograms, the lightest colour curve represents the drift between  $t_0 = 1987$  and  $t = 2006$  and the darkest one is the drift between  $t_0 = 1987$  and  $t = 1988$ .

A first crucial property is the *directed* aspect of the drifts: when the words progressively drift away from their initial representation in a directed fashion. On 10 % of the dataset, the DBE model shows well this behaviour, with a very clear colour gradient. It is also the case for the other models on this subset. With 1 % of the dataset on the contrary, the ISG model is unable to display a directed be-

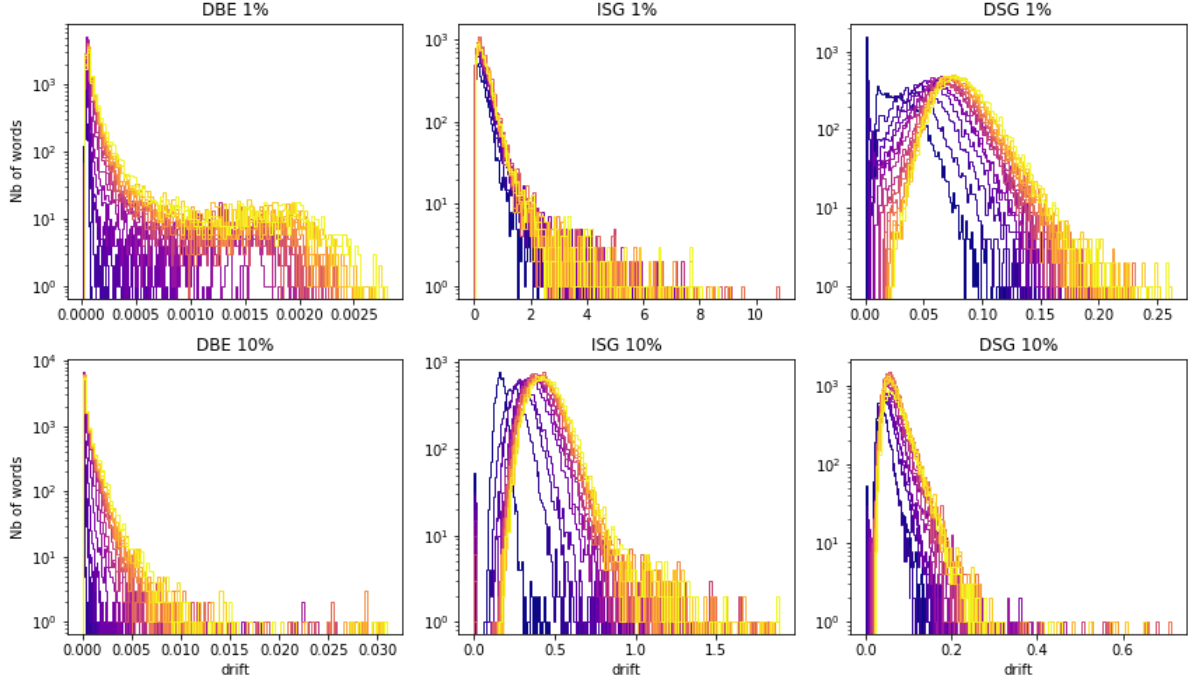


Figure 2: Histogram of word drift for each model on two subsets of the NYT corpus. The drifts are computed from  $t_0 = 1987$  to each successive time step, and superposed on the histogram. The lightest colours indicate drifts calculated until the most recent time steps. The number of words are on logarithmic scale.

haviour (no colour gradient), while the two other models do. This is justified by the use of the diffusion process to link the time steps in equations 2 and 5: it allows the DSG and DBE models to emphasise the directed fashion of drifts even in the situation of scarce data.

The second property to highlight is the capacity of the models to discriminate words that drift from words that stay stable. From the human point of view, a majority of words has a stable meaning (Gulordava and Baroni, 2011); especially on a dataset covering only two decades like the NYT. The DBE model has a regularisation term (equation 6) to enforce this property, and a majority of words have a very low drift on the histogram. However, on 1 % of the dataset, this model can not discriminate very high drifts from the rest. The ISG and DSG models have a different distribution shape, with the peak having a drift superior to zero. The only words that do not drift on their histograms are the one that are absent from a time step.

To conclude, both the DBE and DSG model are able to detect directed drifts even in the 1 % subset of the NYT corpus, while the ISG can not. However, the drift distributions of the DBE and DSG models have a much shorter shorter tail on the 1 %

subset than on the 10 % subset: they are not able to discriminate very high drifts from the rest of the words in extreme scarcity situation.

#### 4.4 Regularisation Attempt

To tackle the weakness of the DBE and DSG models on the smallest subset, we attempt to regularise their loss in order to control the weight of the highest and lowest drifts. Our goal is to allow the model to:

- better discriminate very high drifts;
- be less sensitive to noise, giving lower weight to very low embedding drifts.

We test several possible regularisation terms to be added to the loss. The best result is obtained with the *Hardshrink* activation function, which is defined this way :

$$\begin{aligned} \text{HardShrink}(x) &= x, \text{ if } x > \beta \\ &= -x, \text{ if } x < -\beta \\ &= 0, \text{ otherwise} \end{aligned} \quad (8)$$

For the DSG and DBE models, we add to the loss the following regularisation term, amounting to a thresholding function applied to the drift:

$$\text{reg}_\beta = \alpha * \text{HardShrink}(\text{drift}(U_t), \beta) \quad (9)$$

Where  $\alpha$  is the regularisation constant to be tuned,  $\beta$  is the threshold of the hardshrink function, and the drift is computed according to equation 7. The regularisation term is minimised. The activation function acts as a threshold to limit the amount of words having an important drift. We choose  $\beta$  as the mean drift for both models.

For both DSG and DBE, the right tail of the distribution of the drifts with regularisation (Figure 3) is much longer than in the original model (Figure 2). Moreover, in the case of the DSG model, more words have a drift very close to zero.

To conclude, the regularised DSG model considers more words as temporally stable. Furthermore, regularising the loss of the dynamic models allows to better discriminate extreme word embedding drifts for very small corpora.

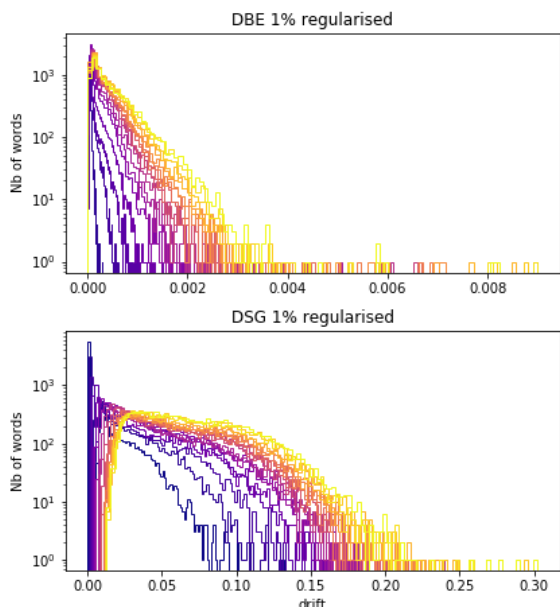


Figure 3: Histogram of word drift for the DBE and DSG regularised models on the 1 % subset.

## 5 Summary & Future Work

To summarise, we reviewed three algorithms for time-varying word embeddings: the incremental updating of a skip-gram with negative sampling (SGNS) from Kim et al. (2014) (ISG), the dynamic filtering applied to a Bayesian SGNS from Bamler and Mandt (2017) (DSG), and the dynamic Bernoulli embeddings model from Rudolph and Blei (2018) (DSG), a probabilistic version of the CBOW.

We proposed two initialisation schemes: the internal initialisation, more suited for low volume of data, and the backward external initialisation, more suited for higher volumes and long periods of temporal study. Then, we compared the distributions of the drifts of the models. We conclude that even in extreme scarcity situations, the DBE and DSG models can highlight directed drifts while the ISG model is too sensitive to noise. Moreover, the DBE model is the best at keeping a majority of the words stable. This property, as long as the ability to detect directed drift, are two important properties of a diachronic model. However, both have low ability to discriminate the highest drifts on a very small dataset. Thus, we added a regularisation term to their loss using the *Hardshrink* activation function, successfully getting longer distribution tails for the drifts.

An important future work is the multi-sense aspect of words. Polysemy is a crucial topic when dealing with diachronic word embeddings, as the change in usage of a word can reflect various changes in its meaning. However, the more different senses are taken into account, the more data is needed to tackle it. An evolution of the DSG model presented in this paper to adapt to this problem would be to represent words while taking into account the context of each occurrence of a word to disambiguate its meaning. Brazinskas et al. (2018) propose such model in a static fashion, where word vectors depends on the context and are drawn at token level from a word-specific prior distribution. The framework is similar to the Bayesian skip-gram model from Barkan (2015) used in the DSG model; but the goal is to predict a distribution of meanings given a context for each word occurrence. We are working on adapting this model to a dynamic framework.

## References

- Jean Aitchison. 2001. Language change: Progress or decay? In *Cambridge Approaches to Linguistics*. Cambridge University Press.
- Robert Bamler and Stephan Mandt. 2017. [Dynamic word embeddings](#). In *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. PMLR.
- Oren Barkan. 2015. Bayesian neural word embedding.



- David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670.
- Arthur Brazinskas, Serhii Havrylov, and Ivan Titov. 2018. [Embedding words as distributions with a bayesian skip-gram model](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1775–1789. Association for Computational Linguistics.
- Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. [Outta control: Laws of semantic change and inherent biases in word representation models](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145. Association for Computational Linguistics.
- Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. [Retrofitting word vectors to semantic lexicons](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615. Association for Computational Linguistics.
- Kristina Gulordava and Marco Baroni. 2011. [A distributional similarity approach to the detection of semantic change in the google books ngram corpus](#). In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [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. Association for Computational Linguistics.
- Michael I. Jordan, Zoubin Ghahramani, and et al. 1999. An introduction to variational methods for graphical models. In *MACHINE LEARNING*, pages 183–233. MIT Press.
- Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. [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. Association for Computational Linguistics.
- Kanako Komiya and Hiroyuki Shinnou. 2018. [Investigating effective parameters for fine-tuning of word embeddings using only a small corpus](#). In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 60–67. Melbourne. Association for Computational Linguistics.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. [Statistically significant detection of linguistic change](#). In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. [Diachronic word embeddings and semantic shifts: a survey](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397. Association for Computational Linguistics.
- Matthieu Labeau, Kevin Löser, and Alexandre Alauzen. 2015. Non-lexical neural architecture for fine-grained pos tagging. pages 232–237, Lisbon, Portugal. Association for Computational Linguistics.
- Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. [Investigating Different Syntactic Context Types and Context Representations for Learning Word Embeddings](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2421.
- Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.
- Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.
- Evan Sandhaus. 2008. The new york times annotated corpus. In *Philadelphia : Linguistic Data Consortium*. Vol. 6, No. 12.
- Cicero D. Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826. JMLR Workshop and Conference Proceedings.
- Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in vkontakte social network. In *ICWSM*.

- Terrence Szymanski. 2017. [Temporal word analogies: Identifying lexical replacement with diachronic word embeddings](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453. Association for Computational Linguistics.
- Nina Tahmasebi. 2018. A study on word2vec on a historical swedish newspaper corpus. In *DHN*.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, 1811.06278.
- Xuri Tang. 2018. [A state-of-the-art of semantic change computation](#). *Natural Language Engineering*, 24(5):649–676.