Building Brownian Bridges to Learn Dynamic Author Representations from Texts

Enzo Terreau, Julien Velcin

Building Brownian Bridges to Learn Dynamic Author Representations from Texts

Enzo Terreau[1] and Julien Velcin[1][0000−0002−2262−045X]

[1]Université de Lyon, Lyon 2, ERIC UR 3083, Bron, France
enzo.terreau@univ-lyon2.fr
julien.velcin@univ-lyon2.fr

Abstract. Authors writing habits fluctuate throughout their lives. This evolution may stem from engaging in new topics, new genres or by the variation of their writing style. However, most representation models aiming at building meaningful authors embedding focus on static representations. They skip the precious time information useful to build more powerful and versatile representations. Only a limited number of methods learn dynamic representations, each dedicated to a time bin. Here we propose a new representation learning model called BARL (Brownian Bridges for Author Representation Learning). BARL uses Brownian Bridges, a Gaussian process, to embed authors as continuous trajectories through time. Leveraging the Variational Information Bottleneck (VIB) framework, it integrates a pre-trained temporal text encoder to encode authors and documents into the same space, learning a distinct dynamic for each author along with a customized variance. We evaluate BARL on several tasks: authorship attribution, document dating and author classification on two datasets from the literature. BARL outperforms baselines and existing dynamic author embedding models while learning a continuous temporal representation space.

Keywords: Author Representation · Dynamic Author Embedding · Variational Information Bottleneck · Brownian Bridges

1 Introduction

Significant work has been made in learning to encode textual information, spanning from representations at subword levels to sentences and documents of various lengths. This encompasses classic methods developed in information retrieval, such as the TF and TFxIDF vector representation, to more recent techniques involving word embedding and contextual representations based on Transformer-based architectures [6, 8]. However, less attention has been dedicated to working at the author level, problem which introduces new features such as topic or style [10]. These representations can be used to solve several downstream tasks, such as author classification or identification, link prediction, and in recommendation systems. Building time-sensitive representations is a must to solve these tasks in a better way.

In this paper, we propose a novel method to construct a latent space that captures the dynamic of authors over time. While previous works have addressed this problem before [7, 9], they do not explicitly build a unique space applicable not only to one but to several sub-tasks. The latent space we aim to build is shown in Fig.1, showcasing the trajectories of several authors through the evolution of their publications over time.

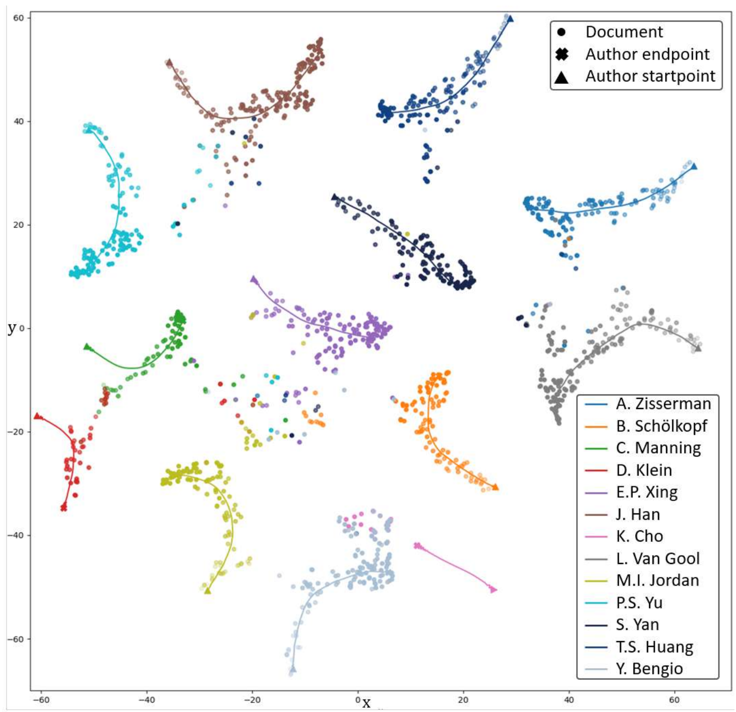

Fig. 1. 2D T-SNE projection of selected author trajectories from the S2G corpus. The color gradient corresponds to the publication date to observe the evolution of its publication over time. Each trajectory goes from its start point ▲ to its end point ✖

Our new method, named BARL (Brownian Bridges for Author Representation Learning), is a representation learning method that uses the concept of Brownian Bridges (BB), a Gaussian process previously used for problems such as video features disentangling [5]. BB is a continuous stochastic process that models the transition between two points in the space (start and end points, illustrated by the triangle and cross symbols in Fig.1).

While BB have already been used for dealing with textual content, it has been done at the document level only to increase consistency of long text generation [19]. In summary, the contribution of this paper is threefold:

– We are the first to use BB to address the dynamic of author representations.
– We introduce the novel BARL model in which we adapt the Variational Information Bottleneck (VIB) framework to deal with author dynamics through the BB mechanism.
– We demonstrate the effectiveness of BARL through quantitative and qualitative experiments, comparing it to the existing literature.

Section 2 delves deeper into related work while section 3 provides details on how we setup this new model BARL by inserting BB into a classic VIB framework. Section 4 showcases the effectiveness of BARL by leveraging the latent space to address several quantitative tasks: author identification, date estimation, and author classification. After a qualitative analysis and the representation space, we present our conclusions in Section 5.

2 Related Works

Leveraging temporal information is crucial for solving today's NLP tasks, as demonstrated by [12] in the context of language models. This has primarily been addressed at the word level [4] and at the sentence level by employing a simple mechanism, such as using the [CLS] token to represent the whole text [2]. Some models have been used to predict the date of a given document, such as NeuralDater [18] and TempoBERT [15]. In this paper, we use TempoBERT as a module in our architecture, although any temporal-oriented model can be considered as alternatives. TempoBERT specializes BERT on the masked word prediction task by adding specific tokens indicating the date of sentence creation.

Constructing author representations over time has been scarcely explored. Early methods date back to the 2000s [16]. In these works, time is split into discrete bins, and word and author representations are estimated for each bin. Following the same temporal discretization, a more recent work, Dynamic Author Representation (DAR), uses an LSTM to model the temporal evolution of author representations [7]. In this model, words, documents and authors are not embedded in the same latent space. Similarly, Dynamic Gaussian Embedding of Authors (DGEA) assumes that documents are drawn from a Gaussian distribution depending on authors. Two implementations, one based on Kalman filters and one based on a deep learning architecture, are tested, but as DAR they both rely on a discretization of the time span. Our model, BARL, stands out as the only one capable of embedding both documents and authors into a unique continuous latent space. This feature is a crucial prerequisite for capturing author trajectories over time. This constitutes the main contribution of our work.

3 BARL: Brownian Bridges for Author Representation Learning

3.1 Background

We note D the set of all documents and A the set of authors. In this work, we assume that each document $d_a^t \in D$ is written by one author $a \in A$ at time

$t \in \mathbb{R}^+$. We restrict t to an interval that corresponds to the first timestamp (set to 0, by default) and the maximum timestamp $T \in \mathbb{R}^+$, so $t \in [0, T]$. Our objective is then to build a latent space that includes both document representations z_t^a (i.e., the latent representation of the document d_a^t) and author representations h_t^a (which can differ from z_t^a because there is an interpolation, see below). We assume that there is only one document written by a at time t. This is a mild assumption because the scale of t can be adjusted accordingly.

3.2 Using the Brownian Bridges

In BARL, we use the Brownian Bridge (BB) to build author trajectories in the latent space. In this framework, the trajectory of one author is fully characterized by its initial coordinate $h_0^a \in \mathbb{R}^r$ and its terminal coordinate $h_T^a \in \mathbb{R}^r$, which can both be noted by $H^a = (h_0^a, h_T^a)$, where r is the dimension of the latent space. All the intermediate vector representations h_t^a are interpolated between these two points, assuming a Gaussian noise that increases with the temporal distance from the two endpoints:

$$p(h_t^a | h_0^a, h_T^a) = \mathcal{N}\Big((1 - \frac{t}{T})h_0^a + \frac{t}{T}h_T^a, \frac{t(T-t)}{T}\Big) \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with μ mean and σ^2 variance with diagonal given in Eq. 1.

In the following, we aim to make latent representations of documents z_t^a close to their related author representation h_t^a. This mechanism, inspired by the Time Control model of [19], is illustrated in Fig.2.

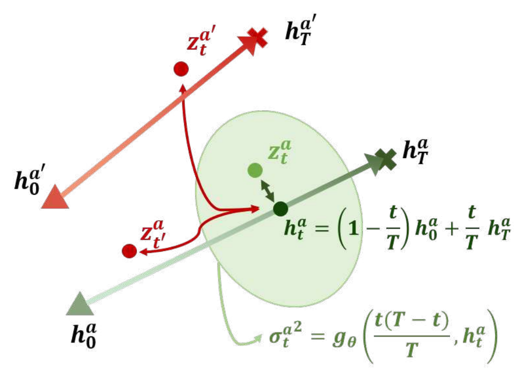

Fig. 2. Illustration of two authors' trajectories (a and a'). We would like to make z_t^a, in green, close to h_t^a, while we make documents farther from the other authors (such as $z_t^{a'}$) or documents written at another time (such as $z_{t'}^a$), in red. h_t^a is the interpolation between the start point h_0^a (▲) and the end point h_T^a (✖)

3.3 Variational Information Bottleneck

In the Time Control theory [19], the initial and terminal states are fixed for all documents, which is not compatible with our aim that is to *learn* the author representation. This leads us to propose a new model that integrates the BB mechanism into the VIB framework, which precisely aims to learn the latent representations z (in our case, both H^a and z_t^a). The general objectif of VIB is the following [1]:

$$\arg\max_z I(z,y) - \beta.I(z,x) \tag{2}$$

where (x, y) are the input and output data (y can be the labels in a classification problem), z is the latent representation, $\beta \in \mathbb{R}^+$ is the trade-off hyper-parameter and I is the Mutual Information. VIB is based on the Information Bottleneck principle [17] that aims at maximally compressing the information in z, such that z is highly informative regarding the labels (i.e., z can be used to predict the labels y). It boils down to minimizing the following loss using the lower bound introduced in [1]:

$$\mathcal{L}_{vib} = -\mathbb{E}[\log q(y|z)] + \beta.KL(p(z|x)||q(z)) \tag{3}$$

where $\mathbb{E}$ is the usual expectation, $q(y|z)$ is the variational approximation of $p(y|z)$, $q(z)$ is for $p(z)$ and $KL(\cdot||\cdot)$ stands for the Kullback–Leibler divergence.

3.4 Learning Author Representations

In our context, we have naturally built author/document (z_t^a, z_t^d) pairs as positive examples (ie., $y = 1$), which are complemented by sampling negative pairs (ie., $y = 0$): $(z_{t'}^a, z_t^d)$ where the time differs, $(z_t^{a'}, z_t^d)$ where author a' differs, or both. In this work, we propose a new loss function that is adapted to our problem:

$$\begin{aligned}\mathcal{L}_{BARL} = &-\mathbb{E}_{p(z_t^a|d_t^a),p(h_t^a|a)}[\log q(y|z_t^a, h_t^a)] \\ &+ \beta[KL(p(h_t^a|a)||q(h_t^a)) + KL(p(z_t^a|d_t^a)||q(z_t^a))]\end{aligned} \tag{4}$$

where the probability of a label y (0 for negative pairs and 1 for positive pairs, see above) is given by [13]:

$$q(y = 1|z_t^a, z_t^d) = \sigma(-c_a||z_t^a - z_t^d||_2 + e_a) \tag{5}$$

where σ is the sigmoid function, $c_a \in \mathbb{R}_*^+$ and $e_a \in \mathbb{R}$ are learnable parameters. Expectation in Eq.4 is intractable for most deep encoders. However, we can approximate it by sampling L observations by training example using the reparameterization trick of VAE [11]:

$$z_t^a = \mu_t^a + \eta_t^a \odot \epsilon,\ h_t^a = (1 - \frac{t}{T})h_0^a + \frac{t}{T}h_T^a + \frac{t(T-t)}{T}\epsilon \text{ with } \epsilon \sim \mathcal{N}(0,1) \tag{6}$$

where z_t^a is fully determined by its mean $\mu_t^a = f_\theta(d_t^a)$ and variance $\eta_t^{a2} = g_\Theta(d_t^a)$, f_θ and g_Θ two functions being learnt as in standard VAE. We can point out that the VIB framework perfectly fits the Gaussian process of BB.

3.5 Model Architecture of BARL

A schematic representation of our model is proposed in Fig.3.

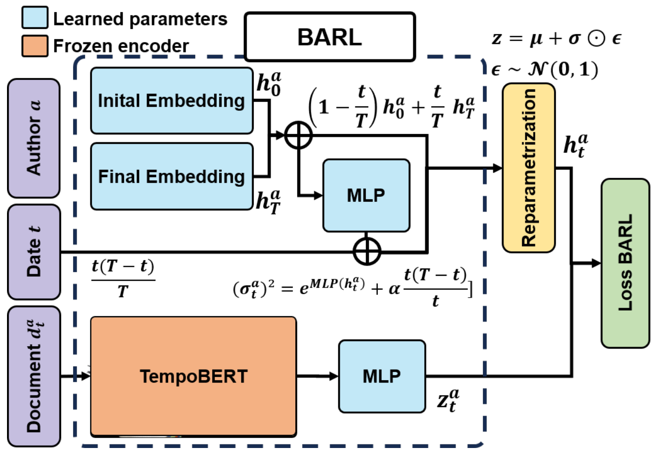

Fig. 3. Architecture of our model **BARL**. At the learning stage, we provide as input the document along a timestamp (a different timestamps for negative examples) and author (a different author for negative examples). During the inference, we can encode any unseen document and interpolate known author representation for any date.

For each author we learn their start and end points (h_0^a and h_T^a) through two embedding layers. Each point matches each author first and last work date. In Section 4 we evaluate the extrapolation capacity of BARL by setting a maximum date T greater than the last known writing date of every author.

In the BB modeling, the variance is only time-based (${\sigma_t^a}^2 = \frac{t(T-t)}{T}$). Each author has its own writing dynamic, with drastic topic or stylistic changes, thus following [9] we compute a log-variance using a 2-layers MLP with `tanh` and `LeakyReLU` activation functions on the author representation h_t^a. Our final variance is given by: $(\sigma_t^a)^2 = e^{\mathrm{MLP}(h_t^a)} + \alpha\frac{t(T-t)}{T}$. α is a learnable parameter we add which weights the relative importance of both variance, initialized to one.

The entering block of our model for documents maps a document in natural language to a vector. We propose to use a pre-trained text encoder. Models that are pre-trained on large datasets are now easily available online. They have been proved successful on many NLP tasks with a simple fine-tuning phase. The VIB framework allows to naturally introduce a pre-trained text encoder. Here we use a frozen TempoBERT [15], a BERT model fine-tuned on temporal masking, and a 3-layers MLP with `LeakyReLU` activation. We do not compute any document variance as experiments give better results without it. Note that any text encoder (ideally time-based) can be used and even fine-tuned during the training stage, making our model language agnostic to the very choice of the encoder.

4 Experiments with BARL

4.1 Datasets

We evaluate our model on two datasets of short documents prepared by [7] and used in the dynamic author embedding literature. Basic statistics are summarized in Table 1.

New York Times (NYT) The NYT corpus was introduced by [20] and is a set of New York Times article headlines from 1990 to 2015. It comes from various sections, from sports to politics. For each title we have its unique author and its full publication date.

We perform a stratified split by authors into train, test and validation sets with a 70/20/10 ratio. We use it in an imputation scheme to evaluate the modeling capability of BARL, each set coming from the whole time span.

Semantic Scholar (S2G) The S2G corpus was introduced by [3]. It is a set of scientific article titles published in machine learning conference from 1985 to 2017. For each article we only have its year publication date (corresponding to one timestep) and it may have multiple authors.

Each distinct author of a document yields a unique (author, document) pair. We perform the same imputation scheme as for NYT but we also use it to evaluate the ability of BARL to extrapolate. For each author we split into train, test and validation sets with a 70/20/10 ratio following the chronological order. Thus, in the prediction scheme the last timesteps (which differs for every author) are not seen during the training stage.

Table 1. Basic statistics of both datasets NYT and S2G

Dataset	Authors	Avg. Tokens	Texts by Author	Period
NYT	546	8.4(±2.5)	76(±51)	$[1990, 2015]$
S2G	1117	8.7(±2.9)	48(±27)	$[1985, 2017]$

4.2 Parameter Settings and Competitors

We select hyper-parameters using a grid search on each validation set. We use a $5e^{-4}$ learning rate with linear decay for 100 epochs and apply early stopping. To fasten the learning stage, we add a writing year token at the beginning of each training example for the first 5 epochs, with an increasing masking probability. We also compare BARL to its variations with only a time-based variance (BARL_t), without any variance (BARL_{novar}) and with an L_2 loss rather than the VIB framework (BARL_{L_2}). The results of these ablation studies are discussed in a dedicated section (see Section 4.6). Our code is publicly available[1].

[1] https://github.com/EnzoFleur/barl

We compare our model against several baselines. First, we consider the Universal Sentence Encoder (USE) [6], a very powerful sentence encoder. Then, BERT models [8] fine-tuned on the authorship attribution task (BERT$_A$), on the document dating task (BERT$_T$) and on both tasks (BERT$_{T+A}$). We also evaluate TempoBERT [15] alone. Finally, we compare BARL with two recent dynamic author embedding methods detailed in Section 2: DAR[2] (Dynamic Author Representation) [7] and DGEA (Dynamic Gaussian Embedding of Authors) [9]. We choose the K-DGEA version of DGEA as it is the fastest and obtain similar results than R-DGEA. For both models we use the configuration given by the authors as they were both trained on the same datasets. We evaluate our model on three tasks, with 10 repetitions for each competitors corresponding to the standard deviation in each result table.

4.3 Results in Authorship Attribution

Authorship attribution consists in assigning to each document its author(s). We use accuracy and coverage error (CE) as metrics. For S2G, we use the usual metric for multi-author corpus: Label Ranking Average Precision (LRAP). For each document embedding we choose its closest author embeddings using cosine similarity. Results are shown in Table 2. Our model BARL ranks first among author embedding methods and third overall (second on S2G in prediction), with at least a two-points increase in coverage error against K-DGEA. DAR cannot produce meaningful document embedding without the author information which is fatal here. BERT$_A$ performs the best even if USE is better in coverage error on NYT. USE confirms its huge modeling capability as an on-the-shelf sentence encoder, challenging fine-tuned models.

Authorship attribution is harder on S2G than on NYT, with more authors and a more specific vocabulary. The time information is more important in a context of scientific publication: this explains why TempoBERT and BERT$_{T+A}$ both achieve better ranks on S2G. Our model is competitive with fine-tuned encoder on a unique specific task while its objectives are multiple: grasp the link between an author and its production with its evolution through time.

4.4 Results in Document Dating

Document dating aims to predict the publication year of each document. The two associated metrics are accuracy and mean absolute error (MAE). We use a K-Nearest-Neighbors as classifier to predict each date, except for baselines with a classification head and TempoBERT. This task is not suitable for prediction as we cannot predict unseen years. Results are shown in Table 3.

Our model BARL achieves the best results on every axis, except MAE on S2G. BARL outperforms fine-tuned model on the document dating task. BERT$_{T+A}$ gets better results than BERT$_T$ which shows the contribution of the author information to predict a document writing period. BARL seems to make

[2] https://github.com/edouardelasalles/dar

Table 2. Authorship attribution on NYT in imputation and on S2G in imputation and prediction. Best models in bold, second underlined, std in parentheses.

	NYT (546 authors)		S2G (1117 authors)			
	Imputation		Imputation		Prediction	
Méthodes	**CE ↓**	**Accuracy ↑**	**CE ↓**	**Accuracy ↑**	**CE ↓**	**Accuracy ↑**
USE	**11.1 (0.0)**	12.7 (0.0)	19.9 (0.0)	10.4 (0.0)	22.3 (0.0)	7.6 (0.0)
BERT$_A$	11.4 (0.8)	**14.3 (0.4)**	**12.8 (0.8)**	**13.4 (0.7)**	**17.9 (1.3)**	**8.9 (0.8)**
BERT$_T$	88.2 (1.0)	0.5 (0.2)	90.2 (1.9)	0.8 (0.2)	94.0 (2.3)	1.1 (0.2)
BERT$_{T+A}$	33.4 (0.8)	3.7 (0.6)	15.3 (1.2)	10.6 (0.7)	25.9 (1.2)	6.7 (0.5)
TempoBERT	12.1 (1.1)	10.2 (0.9)	15.9 (1.5)	10.5 (1.1)	21.6 (0.7)	6.9 (0.4)
DAR	47.1 (1.1)	1.1 (0.2)	38.7 (2.0)	1.2 (0.4)	41.8 (1.2)	0.8 (0.4)
K-DGEA	14.5 (1.2)	9.8 (0.8)	20.4 (1.3)	10.1 (0.8)	28.2 (1.1)	6.6 (0.5)
BARL	11.9 (0.8)	12.2 (1.0)	15.5 (1.4)	10.4 (1.0)	20.6 (1.1)	7.3 (0.8)
BARL$_{L_2}$	27.2 (1.1)	10.1 (0.9)	24.2 (1.2)	9.7 (1.0)	35.5 (1.3)	4.3 (0.9)
BARL$_t$	12.5 (0.7)	11.9 (1.1)	15.9 (1.6)	10.3 (1.2)	21.1 (1.2)	7.1 (0.8)
BARL$_{novar}$	13.1 (0.4)	11.3 (0.6)	19.9 (1.1)	10.1 (0.6)	22.3 (1.1)	7.0 (0.9)

the best of it. Topics and vocabulary in S2G is more time-related as the average MAEs are smaller, even with a bigger global time span.

The two specialized author embedding models are far behind BARL, as they use a discrete representation of time which allows less smoothness and precision in dynamic document representation. The results in both tasks, authorship attribution and document dating, confirm that the main objective of our model has been reached.

4.5 Results in Author Classification

The last task focuses is author classification on S2G corpus in prediction. We associate each author a to the conference (IJCAI, ACL, EMNLP, ...) in which he or she published the most at each timestep $(t_0^a, t_1^a, ..., t_k^a)$ of the training set. We aim to predict their conference for the last timesteps $(t_{k+1}^a, ..., T^a)$. Here we only evaluate the author embedding method in their ability to extrapolate authors' dynamics. We use a linear SVM optimized with grid-search as classifier and accuracy as metric.

Results are given in Table 4. In prediction, we aim at evaluating the extrapolation ability of representation models for unseen future timesteps. As DAR produces static and dynamic representations, we show the results for both representations and their concatenation. Here, author embedding models easily outperform the two language models USE and TempoBERT. Even if S2G only provides the publication year, cancelling one of the main asset of BARL, our model shares the first rank with DAR. Most of the information in DAR representations are hold in the static representations of each author. Our model is able to interpolate as well as extrapolate authors' dynamics while producing continuous representations.

4.6 Ablation Study

We now discuss the results of BARL with its variations (see Tables 2 and 3, bottom lines). The worst results are obtained with the simple L_2 loss. The variational framework of VIB allows more versatility which is key to build such a complex representation space. The same conclusion goes when we look at BARL variations without any variance. Using only a temporal variance is enough to get sufficient results on both authorship attribution and document dating tasks. But it restrains the ability to grasp each author dynamic. Adding an author-related variance allows to capture for each author a distinct evolution, either sticking to the same topic or going through a lot of different topics during their career.

Table 3. Results in document dating in imputation. Best model in bold, second underlined, std in parentheses.

	NYT (26 years)		S2G (33 years)	
Méthodes	**Accuracy ↑**	**MAE ↓**	**Accuracy ↑**	**MAE ↓**
USE	10.4 (0.0)	6.8 (0.0)	9.81 (0.0)	5.1 (0.0)
BERT$_A$	10.2 (0.1)	6.3 (0.1)	8.9 (0.5)	5.5 (0.3)
BERT$_T$	13.0 (0.1)	5.5 (0.2)	12.2 (0.4)	4.4 (0.2)
BERT$_{T+A}$	12.0 (0.1)	5.4 (0.2)	<u>12.3 (0.3)</u>	**4.2 (0.2)**
TempoBERT	<u>13.1 (0.2)</u>	5.2 (0.1)	12.1 (0.3)	4.7 (0.5)
DAR	5.6 (0.2)	7.5 (0.3)	4.8 (0.4)	10.0 (1.1)
K-DGEA	10.6 (0.4)	6.7 (0.3)	7.4 (0.3)	6.8 (0.5)
BARL	**13.3 (0.3)**	**4.7 (0.3)**	**12.6 (0.5)**	<u>4.3 (0.4)</u>
BARL$_{L_2}$	10.5 (0.2)	5.3 (0.3)	8.9 (0.2)	5.4 (0.3)
BARL$_t$	12.8 (0.4)	<u>5.0 (0.3)</u>	12.2 (0.4)	4.6 (0.2)
BARL$_{novar}$	11.7 (0.3)	5.1 (0.2)	11.8 (0.3)	4.8 (0.3)

4.7 Qualitative Analysis

We propose a qualitative analysis of authors' paths learned by BARL on S2G (Fig.1). We represent the ten most prolific authors and some of their co-authors. Authors that have a short publication time span (K. Cho, 4 years, S. Yan, 13 years) are represented by shorter paths. Our model represents locally each author dynamic. Each point cloud is also more spread at the end of each author trajectory, as the quantity of topics an author works on tends to diversify overs time. It is even more the case for authors with publications in lots of various conferences (e.g., Y. Bengio, P. Yu). Finally, co-authorships clearly emerge with close trajectories (e.g., C. Manning and D. Klein, K. Cho and Y. Bengio). These few examples illustrate the modeling power of BARL and its potential to represent authors evolution through time, which takes the co-authorship features into account.

Table 4. Accuracy on author classification for S2G in prediction. Each headline show the portion of training set used.

Methods	S2G Prédiction Accuracy (22 classes) 100 %	75%	50%	25%
USE	28.8 (1.5)	28.3 (1.7)	25.8 (2.4)	24.8 (1.6)
TempoBERT	29.2 (2.3)	28.6 (2.5)	27.5 (1.6)	26.1 (1.6)
DAR (dynamic)	17.5 (1.3)	17.4 (1.4)	17.4 (2.3)	17.2 (1.5)
DAR (static)	34.5 (1.1)	34.2 (1.8)	34.2 (2.0)	33.4 (1.6)
DAR (concat)	35.3 (1.6)	35.0 (1.4)	**34.9 (0.9)**	**34.7 (1.2)**
K-DGEA	35.0 (2.1)	34.7 (2.1)	34.2 (1.8)	33.0 (1.9)
BARL	**35.7 (2.0)**	**35.5 (2.0)**	34.8 (1.8)	34.5 (1.9)

5 Conclusion

We presented BARL, a dynamic author and document embedding model based on Brownian Bridges to represent authors as continuous trajectories through time. To fit this Gaussian modeling, BARL integrates BB into the VIB framework, which brings more versatility and smoothness when capturing author evolution. Our model outperforms existing works in dynamic author representation and it is competitive with specifically fine-tuned encoders. It encodes document and author into the same space and it can integrate any pre-trained temporal text encoder. In future works, we will incorporate more advanced encoders trained on event extraction to process longer and more complex texts (see [18]). We will also test other Gaussian processes ([5]) and add more trajectory points to grasp more complex evolution. We also plan to use this framework at the document scale to embed each document as a trajectory of sentences and see if specific stories schemes arise for example, following [14].

Acknowledgements This work was granted access to the HPC/AI resources of IDRIS under the allocation 2023-AD011012369R2 made by GENCI. The author(s) acknowledge(s) the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-19-CE38-0007 (project LIFRANUM).

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. Proceedings of the ICLR Conference (2017)
2. Amba Hombaiah, S., Chen, T., Zhang, M., Bendersky, M., Najork, M.: Dynamic language models for continuously evolving content. In: Proceedings of the 27th ACM SIGKDD on Knowledge Discovery & Data Mining. p. 2514 (2021)
3. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., et al.: Construction of the literature graph in semantic scholar. In: Proceedings of the Conference of the ACL: Human Language Technologies. vol. 3, p. 84. NO (2018)

4. Bamler, R., Mandt, S.: Dynamic word embeddings. In: International conference on Machine learning. pp. 380–389. PMLR (2017)
5. Bhagat, S., Uppal, S., Yin, Z., Lim, N.: Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In: European Conference on Computer Vision (ECCV) (2020)
6. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, al.: Universal sentence encoder for english. In: Proceedings of the 2018 Conference on EMNLP: System Demonstrations. pp. 169–174 (2018)
7. Delasalles, E., Lamprier, S., Denoyer, L.: Learning dynamic author representations with temporal language models. In: 2019 IEEE International Conference on Data Mining (ICDM). pp. 120–129. IEEE (2019)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL. pp. 4171–4186 (2019)
9. Gourru, A., Velcin, J., Gravier, C., Jacques, J.: Dynamic gaussian embedding of authors. In: WWW '22. p. 2109–2119. Association for Computing Machinery (2022)
10. Jawahar, G., Ganguly, S., Gupta, M., Varma, V., Pudi, V.: Author2vec: Learning author representations by combining content and link information. In: WWW '16 Companion. pp. 49–50 (04 2016)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. Proceedings of the ICLR Conference (2014)
12. Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., et al.: Mind the gap: Assessing temporal generalization in neural language models. In: NeurIPS. pp. 29348–29363 (2021)
13. Oh, S.J., Murphy, K., Pan, J., Roth, J., Schroff, F., Gallagher, A.: Modeling uncertainty with hedged instance embedding. In: Proceedings of ICLR (2019)
14. Reagan, A.J., Mitchell, L., Kiley, D., Danforth, C.M., Dodds, P.S.: The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science p. 31 (2016)
15. Rosin, G.D., Guy, I., Radinsky, K.: Time masking for temporal language models. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. pp. 833–841 (2022)
16. Sarkar, P., Siddiqi, S.M., Gordon, G.J.: Approximate kalman filters for embedding author-word co-occurrence data over time. In: ICML Workshop on Statistical Network Analysis. pp. 126–139. Springer (2006)
17. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. 37th Allerton Conference on Communication, Control, and Computing p. 368 (1999)
18. Vashishth, S., Dasgupta, S.S., Ray, S.N., Talukdar, P.: Dating Documents using Graph Convolution Networks. In: Proceedings of the 56th Annual Meeting of the ACL. pp. 1605–1615. Association for Computational Linguistics, Australia (2018)
19. Wang, R.E., Durmus, E., Goodman, N., Hashimoto, T.: Language modeling via stochastic processes. In: International Conference on Learning Representations (2022)
20. Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H.: Dynamic word embeddings for evolving semantic discovery. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. p. 673–681. WSDM '18 (2018)