NLP Course Project Report
"Exploring Text Coherence in International Relations Data"

Ali Gorji
Department of
Computer Science
ETH Zurich
agorji@ethz.ch

Alexis Perakis
KOF-ETH
Swiss Economic Institute
ETH Zurich
aperakis@ethz.ch

Evzen Wybitul
Department of
Computer Science
ETH Zurich
ewybitul@ethz.ch

Abstract

Recent approaches have been successful at enforcing long-range coherence among textual documents. We leverage a state of the art framework called Time Control to explore coherence in an international relations setting. We train an encoder network constrained by Brownian bridge dynamics on a newly compiled dataset of conflict event data. We find that despite the promising results on Wikipedia articles, Time Control is less performant on crisis event descriptions. We conclude by suggesting avenues for future improvement.

1 Introduction

Recent work has focused on establishing long-run dependencies in text. Wang et al. (2022) constrain the latent space to enforce long-range time coherence within documents. More specifically, their *Time Control* approach models latent dynamics within a document as fixed and following a Brownian Bridge process. In this project we apply Time Control to the international relations ICBe dataset (Douglass et al., 2022) to explore whether this data displays any underlying coherence.[1] Our work is relevant for researchers studying rules and regularities governing international relations. Uncovering latent coherence in conflict dynamics can for example be used to improve conflict monitoring.

Our explicit research question is the following: "Can we learn from textual crisis descriptions a latent space in which international relations display some notion of coherence?" The project described herein is an empirical assessment of our research question. Our contributions are the following:

- We show that textual descriptions in the ICBe dataset do not contain enough signal for Time Control to reliably learn a Brownian Bridge trajectory.
- At the same time, we note that some general information *is* being learned, based on the comparison with randomized baselines.
- We propose avenues for future research to improve the Time Control framework for ICBe.

[1] We model the underlying coherence with a Brownian bridge stochastic process. Consequently, when we talk about *coherence* in this work, we specifically mean the notion of coherence as enforced by this process.

2 Related Work

Arising from the incompetency of language models to generate globally coherent long text, several attempts have been made to address this shortcoming (Guo et al. (2018); Shao et al. (2019); Bosselut et al. (2018); Kiddon et al. (2016); Puduppully et al. (2019)). One of the leading ideas is to assume an inherent "plan" for a document, either manually designed based on context-specific information (Kiddon et al. (2016); Fan et al. (2019); Hua and Wang (2020); McKeown and Duboue (2001); Stent et al. (2004)), or inferred with minimal assumptions in an unsupervised manner (Oord et al. (2018); Wang et al. (2022)). Oord et al. (2018) encode local latent dynamics using a contrastive representation learning approach, focusing on future prediction using the latent space. Wang et al. (2022) push this idea further by constraining the latent space to follow a specific stochastic process, i.e. a Brownian Bridge process, for a more goal-oriented encoding of the latent dynamics.

3 Data

We use the newly compiled International Crisis Events (ICBe) dataset by Douglass et al. (2022) for our work. This dataset contains both text and tabular data about 471 crises from the 20th and the 21st century. The authors define a crisis as "[...] a change in the type, or an increase in the intensity,

of disruptive interaction with a heightened probability of military hostilities that destabilizes states' relationships or challenges the structure of the international system". For this project, we solely use the text descriptions of each crisis, i.e. a sequence of sentences describing the events in the crisis. These sentence descriptions are historical narratives written by experts using a specific coding scheme. We drop duplicate sentences.

We also replicate experiments by Wang et al. (2022) on the Wikisection (Arnold et al., 2019) dataset for comparison purposes. Wikisection is constituted of Wikipedia articles about cities. These articles are split by sections, each section typically containing a number of sentences. There are a total of 4 sections: "abstract", "history", "geography" and "demographics". These sections appear as tags in the text data only in front of the first sentence of each section.

4 Methods

Our empirical investigation relies on a number of assumptions. First, we posit that there is some underlying coherence in international relations. Second, we assume that this coherence can be extracted from text data, and, more specifically, from the event descriptions in ICBe. Given these assumptions, we attempt to train a model that identifies this coherence in a way that is independent of time of the crisis, geography, belligerent parties etc.

4.1 Modelling Framework

We follow the approach proposed by Wang et al. (2022), where the authors train a Multi-Layer Perceptron (MLP) encoder on top of frozen pretrained GPT-2 (Radford et al., 2019) embeddings. We train our encoder to learn new embeddings under the constraint of time-independent dynamics. These latent dynamics are enforced through a Brownian bridge process. A Brownian bridge process is a pinned Brownian motion process, which can alternatively be seen as a Wiener process with fixed end point. One can also view it as a Gaussian process with a Radial Basis Function (RBF) kernel, where we observe noiseless initial and final points. The density of a Brownian bridge is the following:

$$p(z_t|z_0, z_T) = \mathcal{N}\left(\left(1-\frac{t}{T}\right)z_0 + \frac{t}{T}z_T, \frac{t(T-t)}{T}\right) \quad (1)$$

where z_0 and z_T are arbitrary start ($t = 0$) and end points ($t = T$) respectively. From Equation 1 a Brownian bridge process can be intuitively interpreted as a noisy interpolation between its start and end points.

In our experiments we follow the pinning of Wang et al. (2022) where the first and last sentence in the crisis (and not in the triplet!) have embeddings $z_{first} = \mathbf{0}$ and $|z_{last}| = \mathbf{1}$. While these values are arbitrary, what matters is the process describing the latent dynamics and not the absolute positions of a crisis' start and end point in latent space. For our model to learn the pinning to start and end sentences of a document we use the pinning loss in the spirit of Wang et al. (2022):

$$\mathcal{L}_{\text{pin}} = \text{MSE}(z_{first}, \mathbf{0}) + \text{MSE}(|z_{last}|, \mathbf{1}) \quad (2)$$

where MSE is the mean squared error, z_{first} and z_{last} are the learned embeddings of the first and last sentence in a document respectively.

We train our encoder with a contrastive learning objective. In the contrastive learning paradigm, one is interested in maximizing some notion of similarity between positive observation pairs while minimizing the similarity between negative pairs. As in Wang et al. (2022), we operationalize latent space coherence by defining a contrastive optimization objective for a sentence triplet (x_0, x_t, x_T) with respect to the aforementioned Brownian Bridge process. A sentence triplet (x_0, x_t, x_T) is constituted of an ordered sequence of 3 sentences sampled from a given document (i.e. crisis). The respective triplet observation times $(0, t, T)$ correspond to relative sentence positions in the document, where x_0 and x_T are not necessarily the first and last sentences in the document (i.e. crisis). However, they do correspond to the start and end points of each Brownian bridge process, hence this index convention.

The triplets are loaded in batches, Ω. In turn, we interpret each $(x_0, x_t, x_T) \in \Omega$ as a positive example, while the $x'_t, t' \neq t$ from the remaining triplets $(x'_0, x'_t, x'_T) \in \Omega$ are interpreted as negative examples for the first triplet. We define the contrastive loss contribution l_i for each positive instance sentence triplet i as:

$$l_i = -\log \frac{e^{-\text{d}(x_0, x_t, x_T; f_\theta)}}{e^{-\text{d}(x_0, x_t, x_T; f_\theta)} + \sum_{x_{t'} \in \Omega^-} e^{-\text{d}(x_0, x'_t, x_T; f_\theta)}} \quad (3)$$

where $x_{t'}$ denotes a negative sample and Ω^- the set of negative samples of x_t in the batch considered. We follow the in-batch soft negative sampling as in

Wang et al. (2022). Moreover we use as similarity function the Euclidian distance d of the embedding of x_t to the noisy interpolation in latent space:

$$\begin{aligned}&\mathrm{d}(x_0, x_t, x_T; f_\theta)\\&= \frac{1}{2\sigma^2}\|f_\theta(x_t) - \left(1 - \frac{t}{T}\right) f_\theta(x_0) - \frac{t}{T} f_\theta(x_T)\|\\&= \frac{1}{2\sigma^2}\|z_t - \left(1 - \frac{t}{T}\right) z_0 - \frac{t}{T} z_T\|\end{aligned} \tag{4}$$

where f_θ is our encoder projection into latent space, θ our learned encoder parameters, z_t the learned sentence embedding of x_t and $\sigma = \sqrt{\frac{t(T-t)}{T}}$ the standard deviation of our Brownian Bridge. The full contrastive loss over a batch of size N is then:

$$\mathcal{L}_{con} = \frac{1}{N}\sum_{i=1}^{N} l_i \tag{5}$$

As is the case in Wang et al. (2022), it is important to underline that we do not enforce a single global Brownian bridge process on the whole crisis (i.e. document). Instead, the loss is devised in a way to constrain every triplet of sentences to a separate noisy interpolation. However, we expect the different Brownian bridge processes to align the learned embeddings if the pinning is successful and the effective Brownian bridge variance after training is low.

4.2 Evaluation

The open nature of our task restricts the space of meaningful, universal, quantitative metrics for evaluating the encoder. Ideally, we would set up a number of downstream tasks and benchmarks to analyze the learned embeddings, but that is out of scope for this project. To be able to propose an answer to our research question, we instead first perform a qualitative assessment of the embeddings to see how well they compare to the embeddings from Wang et al. (2022). The strong assumption we are making is that Time Control works on the Wikisection data, as reported.

The second comparison during our evaluation is to visualize whether the model learns to distribute the points in a manner that fits a Brownian bridge process. If they fit this process well, this would suggest that the model discovers an underlying coherent structure of the crises. Under a noiseless interpolation objective, the encoder would attempt to evenly distribute the embeddings along a line, placing the first at **0** and the last (in absolute value) to **1**, depending on the time (order) of the sentences. While this is an asymptotic behaviour, we expect a similar distribution of embeddings in the event of sufficient penalty faced by the model in a Brownian bridge setting.

To assess the ability to generalize we perform the evaluation on a held-out test set, and compare the performance on multiple dataset configurations, including randomized baselines. Namely, for each dataset, we aggregate the embeddings from the test set in each dimension; for an example, see Figure 1. We perform a linear regression to assess the dependency of each dimension on time, and we also bin the embeddings and show the standard deviation for each bin. For comparison, we also plot the "golden" line from 0 to ± 1 (depending on the particular dimension), on which the points should theoretically lay (at least assuming a noiseless interpolation).

We also experiment with a likelihood-based quantitative assessment.

4.3 Correcting Original Code

We found a code inconsistency in terms of the definition of the contrastive loss for Wang et al. (2022). In their framework they use an in-batch soft negative sampling procedure. For a given triplet's (x_0, x_t, x_T) positive sample x_t, the soft negative samples $x_{t'}$ are all the positive instances from the other triplets in the batch. However, for the Brownian bridge process to be defined, any positive or negative instance must be located between the start x_0 and end x_T points given. Thus, what constitutes the negative instance is only the sentence $x_{t'}$ and not its associated time index t', as this time index is inconsistent with the positive triplet's start and end points. In the code of Wang et al. (2022) the noisy interpolation uses the time indexes t' of negative samples to calculate the target embedding instead of the positive time index t. We adapt the loss to avoid situations where the interpolated target embedding is located outside of the Brownian bridge process and is therefore undefined.

5 Experiments

5.1 Ablation Studies

We perform a set of ablation studies to isolate the effect of various hyperparameters on the embeddings of interest. Unless differently specified, the embedding dimension is fixed to 16, for which Wang et al. (2022) observe the best performance

on their downstream tasks.

Dataset We compare the embeddings learned from ICBe sentences and Wikisection sentences to elicit dataset related effects.

Shuffling We compare the embeddings learned on the randomly shuffled sentences of the ICBe dataset to the non-shuffled ICBe dataset. We repeat the comparison between shuffled and non-shuffled sentences for Wikisection. Here there are two distinct shuffling regimes: full shuffling and shuffling of sentences only within document sections.

Latent Dimensions We investigate the effect of embedding dimensionality by experimenting with two different latent space dimensions (2 and 16) on the ICBe data.

Augmentations We investigate the effects of the section tags present in the Wikisection data. We compare the embeddings when trained on sentences with and without the tags.

5.2 Training Details

We use a 4-layer trainable MLP ($d_{in} \rightarrow 128 \rightarrow 128 \rightarrow d_{out}$) on top of the final layer of a frozen pretrained GPT-2 network. We use ReLU as the activation function in our MLP. For training we use SGD with learning rate of 0.0001 and momentum of 0.9. We run our encoder on the Wikisection dataset for 100 epochs and on ICBe for 500 epochs. The difference in epochs is motivated by the size difference between the two datasets.

6 Results & Discussion

6.1 Experiments

Dataset From Figure 1 we can observe that the latent embeddings learned from ICBe do not resemble the Wikisection embeddings. Moreover, we can note that the embeddings learned by Wang et al. (2022) are closer to the noise-free interpolation between the pinned start and end points (the green dotted line). It therefore appears that Time Control, under our current specifications, is not very good at extracting coherence from the ICBe data.

Shuffling Figure 4 shows a striking difference between the shuffled and non-shuffled ICBe dataset. Shuffling the sentences in a crisis leads to constant embedding values. This is in line with our intuition. We do not expect the encoder to be able to learn embeddings that follow a Brownian bridge process as we have artificially broken any sequential structure in the data. Figure 7 also displays an interesting result. It appears that when shuffling the sentences within the same section of the Wikisection dataset, the model is still able to fit a Brownian bridge process. What is more, the embeddings seem to be as good (if not better) than the non-shuffled Wikisection dataset. This leads us to believe that the Time Control framework learns rather long-range (cross-section) coherence. However,further supportive evidence is necessary to confirm this. Shuffling the Wikisection dataset completely leads to embeddings similar to the shuffled ICBe dataset (c.f. Figure 5) [2].

Latent Dimensions Figure 3 is difficult to interpret in terms of quality of Brownian bridge fit. However, the pinning to the end point seems to work better in lower dimensions (i.e. d=2). It is interesting to note that Wang et al. (2022) show a large performance sensitivity to latent space dimensions. From the 3 specifications reported (8,16,32), their 16-dimensional embeddings perform best. This may be indicative of a highly unstable system.

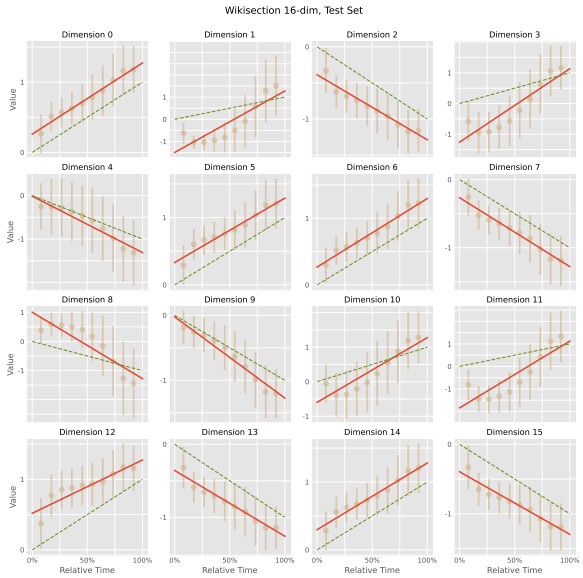
Augmentations Figure 6 displays the embeddings for the Wikisection dataset with and without its section tags. We do not observe a notable difference between the two settings. This result is in line with our intuition, as we do not expect the very small number of augmented sentences (4 per document) to provide the model with significantly more information.

Quantitative Evaluation We do not find that our likelihood-based assessment adds any significant value. For completeness, we include it in the appendix; c.f. Table 1.

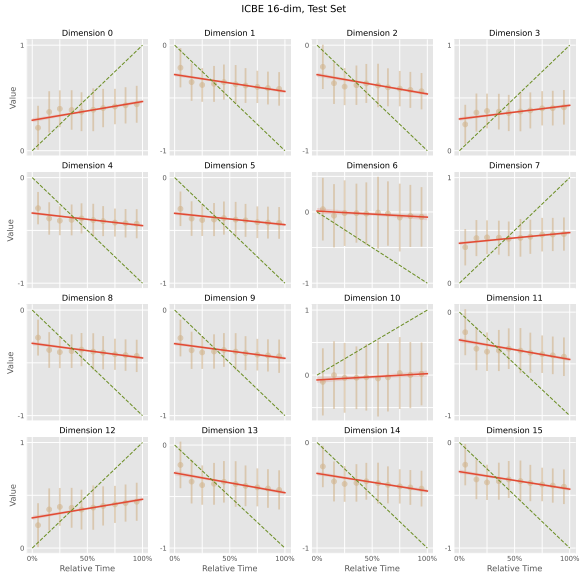
6.2 Dataset

We expect the dataset's nature to explain at least part of the differences observed between the Wikisection and the ICBe results. First of all, our ICBe dataset contains 8312 sentences for a total of 432 documents (i.e. crises), as compared to 108071 sentences and 2823 documents for Wikisection. Moreover, we can posit a large difference in sentence coherence between ICBe and Wikisection. The sentences in Wikisection are naturally more coherent

[2] We vary two parameters by removing the section tags and fully shuffling Wikisection dataset in Figure 5. However, the ablation study results with respect to the section tags permit this conclusion.



(a) Wikisection



(b) ICBe

Figure 1: A comparison of test-set trajectory aggregations on the Wikisection dataset, and on ICBe. Both latent spaces have 16 dimensions.

as they constitute a fully fledged Wikipedia article. In contrast to that, the sentences in ICBe correspond to event descriptions for a particular crisis. They rather resemble standalone sentences than sentences constituting a document. By the nature of ICBe we are therefore working with a smaller dataset that likely contains less coherent sentences, as compared to Wikisection. Still, the ICBe dataset includes a rich hierarchy of event labels that could be leveraged to add informational tags to the crisis descriptions. Further research is needed to verify if these richer and more frequent — compared to the ones in Wikisection — tags can improve the performance of the model.

6.3 Contrastive Objective

Our results from Figure 1 may suggest that the penalty in the contrastive objective faced by the model in areas of high uncertainty is insufficient to induce meaningful training. As can be seen in Equation 1 , the uncertainty is captured in the variance of the Brownian bridge process density. One way to induce more penalty is to reduce the variance tolerated by the Brownian Bridge dynamics. Reducing the variance is equivalent to scaling the Euclidian distance using a temperature parameter τ. This is quite standard in contrastive learning settings and modifies the loss function as follows:

$$-\log \frac{e^{-\mathrm{d}(x_0,x_t,x_T;f_\theta)/\tau}}{e^{-\mathrm{d}(x_0,x_t,x_T;f_\theta)/\tau} + \sum\limits_{x_{t'}\in\Omega^-} e^{-\mathrm{d}(x_0,x_t,x_T;f_\theta)/\tau}} \tag{6}$$

Alternatively, we may attempt improving the signal fed to the model by modifying our negative sampling strategy. Both our setting and Wang et al. (2022) use a contrastive objective where the negative sentence instances are sampled in-batch. In other words, negative samples are sentences taken from other triplets in the batch, thus not necessarily (and most likely not) belonging to the same document (or crisis) as the positive sample. While this approach is efficient and appears to work for Wang et al. (2022), it may be too noisy for our application. Instead, we could benefit from hard negative samples since they carry a stronger signal. The hardest negative sample instance possible for ICBe are sentences from the same crisis as the positive sample, where the negative sample has a time index within 0 and T but different from t. However, discriminating time (or order) at this granularity might be too hard of a task for our model. Slightly less

hard negative samples would be sentences from the same crisis but outside of the triplet time range. Such negative sampling would be in spirit of the Sentence Order Prediction framework used by Lan et al. (2020) to improve coherence. In future work we plan to use both aforementioned adaptations of the contrastive objective to improve our learned embeddings.

7 Conclusion

In this project we explore whether we can use the Time Control framework to uncover the existence of latent dynamics in international relations data. We find it difficult to extract a clear coherent structure from the ICBe dataset event descriptions. Further work is necessary to improve the framework's performance on ICBe. Finally, devising a clear downstream task or quantitative evaluation metrics are essential to better measure the performance of our modelling approach.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184.

Rex W. Douglass, Thomas Leo Scherer, J. Andrés Gannon, Erik Gartzke, Jon Lindsay, Shannon Carcelli, Jonathan Wilkenfeld, David M. Quinn, Catherine Aiken, Jose Miguel Cabezas Navarro, Neil Lund, Egle Murauskaite, and Diana Partridge. 2022. Introducing the icbe dataset: Very high recall and precision event extraction from narratives about international crises. *Computing Research Repository*, arXiv 1906.06349.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Xinyu Hua and Lu Wang. 2020. Pair: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 329–339.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Kathleen McKeown and Pablo A Duboue. 2001. Empirically estimating order constraints for content planning in generation.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3257–3268.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86.

Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. 2022. Language modeling via stochastic processes. In *International Conference on Learning Representations*.

A Supplementary Material

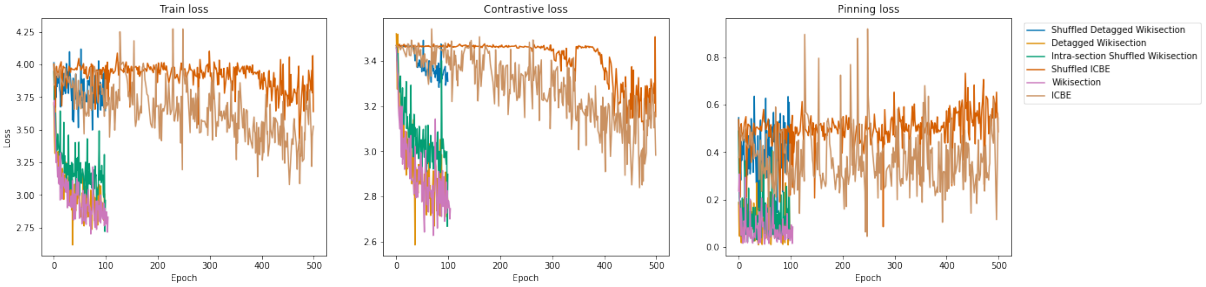
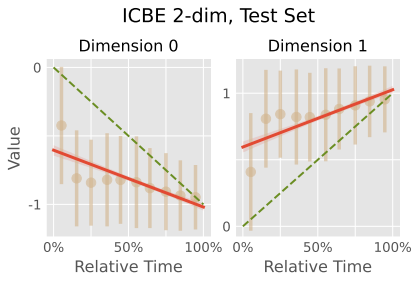


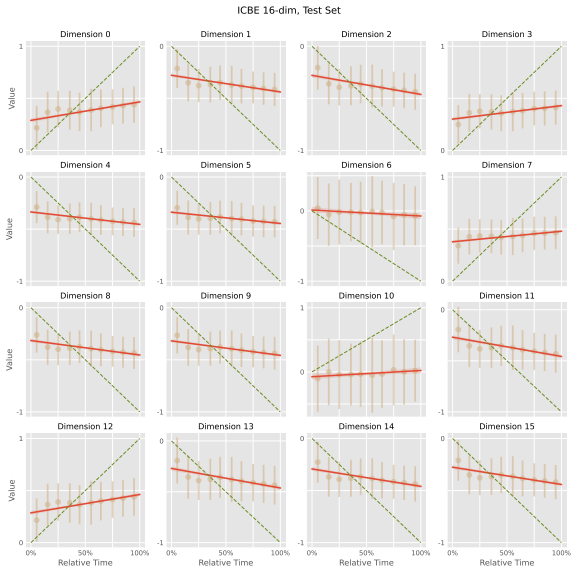
Figure 2: Training losses for the different experiments performed. The curves are not smoothed.

Dataset	Log-likelihood	Simulated log-likelihood
ICBe	-82.41	6.22
ICBe, shuffled	-104.14	
Wikisection, shuffled within section	-357.03	-318.64
Wikisection, detagged	-425.31	
Wikisection	-430.6	
Wikisection, detagged, shuffled	-449.37	

Table 1: Median log-likelihood of a test-set trajectory on different datasets. The likelihood is measured under a Brownian bridge process with max vairance of 0.1, start pinned to $\mathbf{0}$, and end pinned to some vector v, $|v| = \mathbf{1}$, which is chosen separately for each dataset depending on the training process. For comparison, the median log-likelihood of trajectories simulated from the same process is also listed.

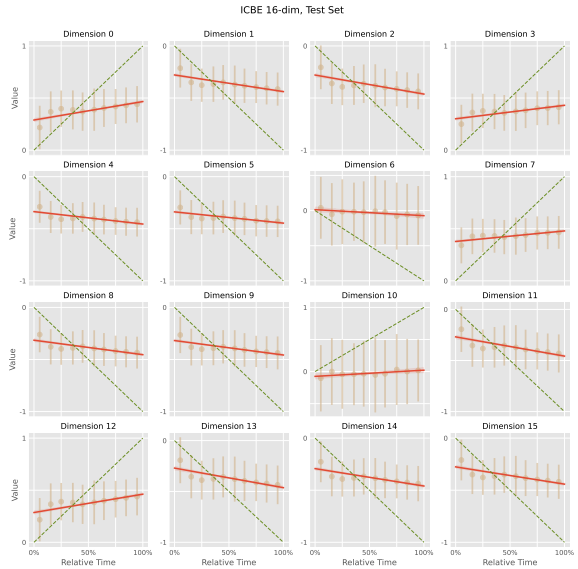


(a) 2-dimensional latent space

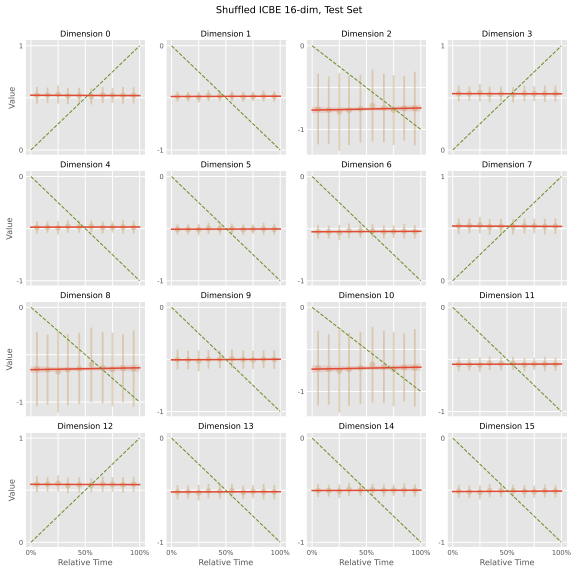


(b) 16-dimensional latent space

Figure 3: A comparison of test-set trajectory aggregations for encoders trained to project into a 2-dimensional and a 16-dimensional latent space.

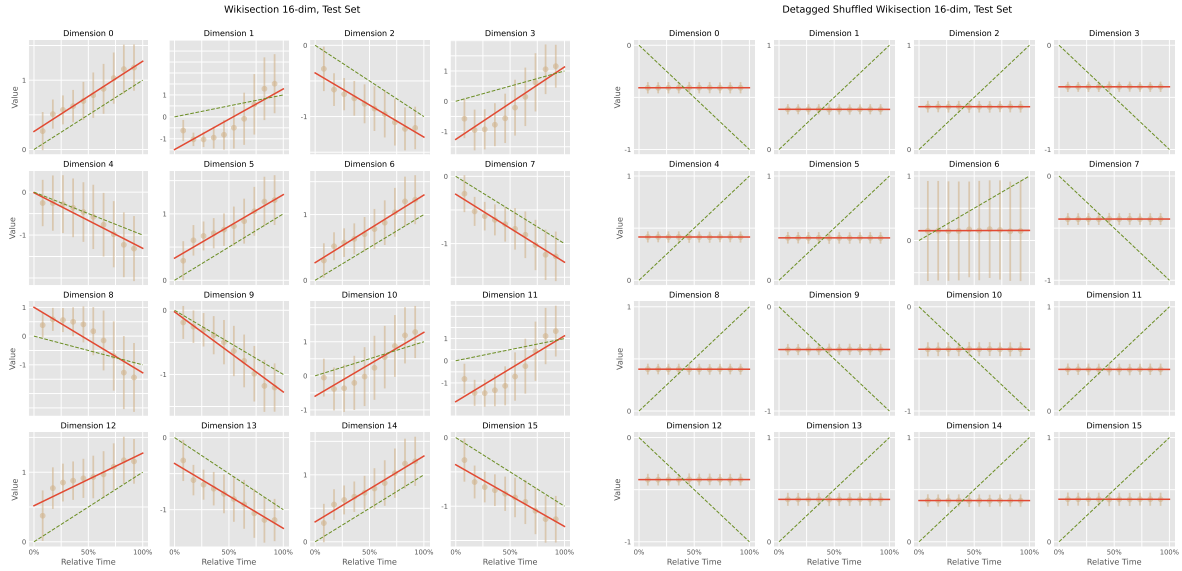


(a) Original ICBe



(b) Shuffled ICBe

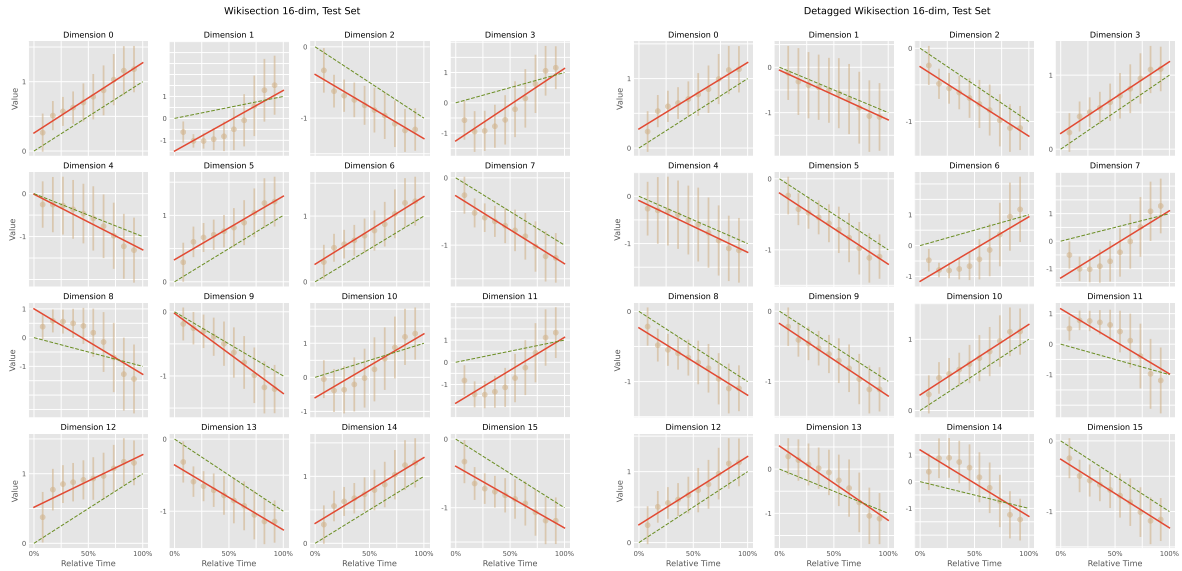
Figure 4: A comparison of test-set trajectory aggregations on ICBe. In (a), the sentences are kept in the original order, while in (b) they are randomly shuffled within each crisis before training.



(a) Original Wikisection

(b) Shuffled Wikisection without section tags

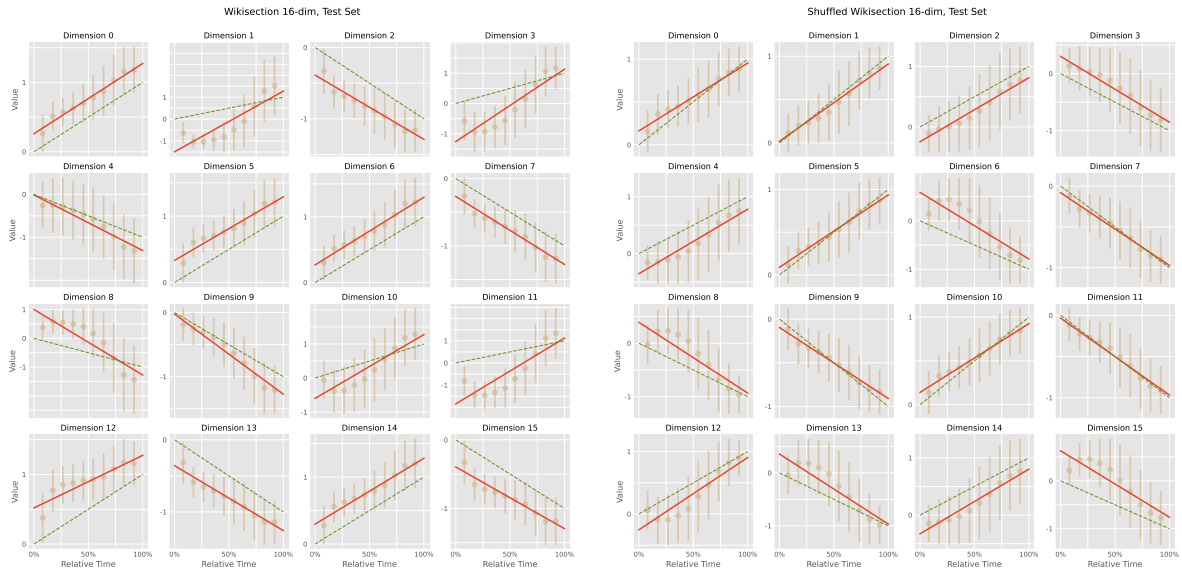
Figure 5: A comparison of test-set trajectory aggregations on Wikisection. In (a), the sentences are kept in the original order, while in (b) they are randomly shuffled within each crisis before training. Additionally, in (b) all section-identifying tags are dropped.



(a) Original Wikisection

(b) Wikisection without section tags

Figure 6: A comparison of test-set trajectory aggregations on Wikisection. In (a), the sentences are kept as-is, while in (b) all section-identifying tags are stripped.



(a) Original Wikisection

(b) Wikisection with sentences shuffled within sections

Figure 7: A comparison of test-set trajectory aggregations on Wikisection. In (a), the sentences are kept as-is, while in (b) the sentences are shuffled within each section. The overall section order is kept constant.