Exploring Topological Invariance in Semantic Embeddings

Fangzhou Gao and Justin Brody
Franklin and Marshall College
PO Box 3003
Lancaster, PA 17604
fgao@fandm.edu justin.brody@fandm.edu

Abstract

We present the result of preliminary explorations of using the topology of embedded manifolds as a semantic invariant. Our main question is whether the topology of large embedded corpora is invariant in the following two senses. First, one might reasonably expect that the same corpus in two languages would give topologically equivalent embeddings. Second, one might reasonably expect that the same corpus embedded by two different embedding models might give topologically equivalent embeddings. In the paper we will justify these intuitions and give preliminary results indicating that they are, to some extent, justified.

1 Introduction

In this paper, we begin an exploration of the possibility of using the tools of topological data analysis to uncover a *semantic signature* for a corpus. Contemporary natural language processing allows for a multitude of ways to embed a corpus into a vector space. A single text can be embedded by multiple models; similarly a single text can exist in multiple languages, and each language can be embedded by multiple models.

For reasons explained in the next section, we hypothesize that the topology of the structure underlying such embeddings provides such a semantic signature. If this hypothesis is correct, then we can use the tools of topological data analysis to uncover the semantic signature of a corpus. In particular, an accurate translation of a text should have the same semantic signature as the original text and persistent homology can be a way to measure translation accuracy.

In our paper we will report on a preliminary exploration of these hypotheses. We will find some promising signs for our hypotheses, but will not fully confirm them.

In particular, we will look at a set of web novels in Mandarin Chinese, along with translations into English and German. These are embedded via a single multi-modal model and we will compare the persistent homology of the resulting manifolds.

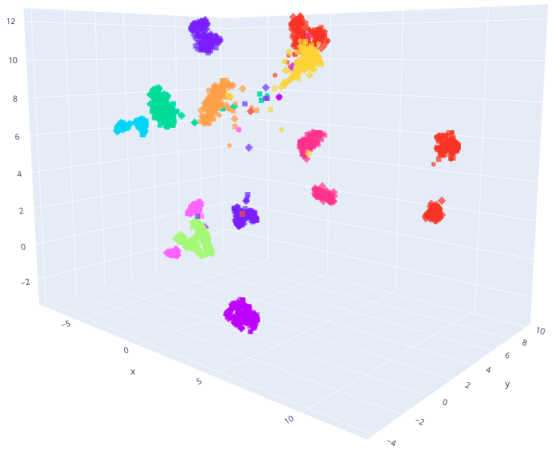

Figure 1: Embeddings of chapters. Regions with similar colors correspond to embeddings of the same chapters in different languages. The shapes of these embeddings are often very similar.

2 Preliminaries

Our fundamental work here is with embeddings of corpora. Transformer neural networks can fundamentally be thought of as embedding architectures. In particular, encoder architectures (like BERT) provide methods of transforming words into vectors in such a way that the embeddings absorb the context of the other words in the sentence, while a special CLS token can be thought of as absorbing the context of the entire sentence. These embedded sentences are our fundamental objects of study.

The intuition behind these embeddings of sentences into vector spaces is that sentences with similar meanings will be embedded close to each other in the vector space (according to some metric). This is the basis of using sentence embeddings for topic modelling; for example this is the basis of BERTopic [4]. Sentences with similar meanings

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 320–324

will be embedded close to each other and form clusters which correspond naturally to topics.

We take this as our foundation – that high quality embedding models will embed sentences with similar meanings close to each other. From just this assumption, what can we say about the embeddings produced by a corpus? We certainly have no right to assume that two different embeddings of the same corpus will have the same geometry [1]. However, we might expect that they will have the same topology, since topological properties are precisely those which are determined from flexible notions of "closeness". In particular, given a particular corpus C and another version of that corpus D (say in a different language), we might expect a high quality embedding model to embed C and D in ways which produce topologically similar shapes. If a model maps sentences to a universal language before embedding them, then we might expect to see much stronger relations between the embeddings of C and D, with identical embeddings being the strongest possibility and embeddings with the same shape being a slightly weaker possibility. We might further expect two different high quality embedding models to embed C in different but topologically similar ways. These observations lead us to conjecture that different embeddings with the same underlying meaning should have the same topological structure.

Measuring topological structure presents a number of difficulties which the tools of *topological data analysis* (TDA) are designed to mitigate [2]. We assume that the embeddings of a corpus form a discrete sample from some underlying continuous manifold, which we term the *semantic manifold*. The tools of TDA are then designed to uncover this manifold and provide a signature of it. In particular, we will use persistent homology to measure the topological structure of the manifolds corresponding to different embeddings of the same corpus. We will then use the bottleneck distance to measure the similarity of the persistent homology of different embeddings.

One of the early challenges in topology was to find a way to characterize the shape of a manifold. It turns out [5] that one way to do this is to essentially look at the holes in a manifold. Intuitively, we can think of circles (and their higher dimensional analogues – spheres) as being filled or unfilled depending on whether or not they can be shrunk down to a point. A canoncial example is a circle that is the boundary of a disk. The disk can be continuously shrunk down to a point without obstruction, so this circle would not describe any kind of unfilled hole. On the other hand, a circle that is the boundary of an annulus cannot be shrunk down to a point, so this circle would describe a unfilled hole. Homology groups are then ways of describing a manifold by giving a canonical count of the number of unfilled holes in various dimensions – counting the number of fundamental [2] unfilled circles (or 1-dimensional spheres) in a manifold M gives $H_1(M)$, the first homology group of M.

Standard homology requires a fixed, continuous shape (a manifold). In our context, we only have a finite sample of discrete data points (a point cloud) and need to infer the underlying manifold. Persistent homology solves this by creating a *filtration of simplicial complexes* from the point cloud. The idea is connect close-by points and try to estimate whether they are connected as part of a line (unfilled), a triangle (filled) or higher dimensional analogue of a triangle. Specifically, for each r we place a sphere of radius r around each point and looking at how these spheres overlap. If two points have overlapping spheres, they are assumed to lie on the same line segment , if three points have overlapping spheres they are assumed to lie on the same triangle, and so on.

When $r = 0$, the filtration simply produces isolated points. As r increases, the spheres begin to overlap and edges will start to appear. As r continues to grow, unfilled holes will form in this complex; this is the "birth" of an unfilled hole. As r grows even larger, those holes will eventually get become filled in by the expanding spheres; this is the "death" of an unfilled hole.

Persistent homology is the study of tracking the birth and death of these topological features (connected components, loops, voids) across all possible scales of r. The main idea is that features that persist over a large range of r represent true, underlying signal, while features that die quickly are topological noise.

[1] Intuitively, that would require that if S, T are similar sentences, then not only will two high quality models embed S and T close to each other, but will also embed them at similar angles from each other.

[2] Fundamental in the sense that no other circle is "homotopy-equivalent" to it. A discussion of this term is beyond the scope of this paper.

3 Methodology

We use the GuoFeng Webnovel corpus [10, 9], a large-scale multilingual collection of web fiction spanning multiple languages and genres. Following the dataset release structure, we use Chinese (ZH) and English (EN) data from V1 and German (DE) from V2. As the DE portion covers fewer books than ZH/EN, all cross-lingual experiments are restricted to the *intersection* of `book_ids` available across the relevant languages, ensuring equal book coverage across pairs. We report intersection sizes for each language pair alongside the corresponding evaluation results.

For EN and ZH, the raw source files contain explicit structural markup for the book and chapter boundaries (e.g., `<BOOK id="...">`, `<CHAPTER id="...">`). We parse these tags to assign `book_id` and `chapter_id` metadata to each line, strip the markup, and remove chapter-title lines (e.g., lines immediately following a `<CHAPTER ...>` tag that contain chapter headings). For ZH, we additionally remove dataset-specific dialogue markers when the marker pattern is detected.

We perform sentence segmentation using **Stanza** [6] with sentence splitting enabled for all three languages. Each segmented sentence is stored alongside its metadata (`lang`, `book_id`, `chapter_id`, and within-line sentence index).

We embed all sentences using a single multilingual model, enabling all three languages to be represented in a shared vector space. We use **BAAI/bge-m3** [3] via the SentenceTransformers library [7] and produce L2-normalized embeddings. The resulting vectors are stored as a matrix $E \in \mathbb{R}^{N \times d}$, where N is the total number of sentences and d is the embedding dimension, aligned with the sentence metadata table.

To obtain a more stable unit of analysis than individual sentences, which are short and noisy in web fiction, we compute chapter embeddings by mean-pooling sentence embeddings within each (`book_id`, `chapter_id`) group:

$$c_{b,k} \;=\; \frac{1}{|S_{b,k}|} \sum_{s \in S_{b,k}} e_s, \tag{1}$$

where $S_{b,k}$ is the set of sentences in chapter k of book b and e_s is the embedding of sentence s. We L2-normalize $c_{b,k}$ after pooling. Chapters with fewer than τ sentences are excluded, where τ is swept as part of a robustness analysis. Each book b in language ℓ is then represented as a point cloud of chapter vectors:

$$X_{b,\ell} \;=\; \{c_{b,k}\}_{k=1}^{K_b}. \tag{2}$$

Although multilingual embedding spaces can preserve cross-lingual structure, aggregated chapter and book embeddings exhibit a strong global *language offset*. To correct for this, we apply language centering by subtracting a per-language centroid:

$$\tilde{x} \;=\; x - \mu_\ell, \qquad \mu_\ell \;=\; \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} x_i, \tag{3}$$

where the sum ranges over all vectors in language ℓ belonging to the common-book intersection, ensuring comparability across languages. Unless otherwise noted, all analyses use centered embeddings.

To isolate the respective contributions of language and content to embedding similarity, we evaluate a 2×2 factorial design at the chapter level, crossing two factors: (i) Language: same (**SL**) vs. different (**DL**) , and (ii) Content: same book (**SC**) vs. different books (**DC**). This produces the four conditions SL–SC, SL–DC, DL–SC, DL–DC.

To characterize higher-order geometric structure in each book's chapter point cloud beyond mean similarity, we compute Vietoris–Rips persistent homology on $X_{b,\ell}$ under the Euclidean metric using **Ripser** [1]. We compute persistence diagrams up to homology dimension 1, capturing connected components (H_0) and loops (H_1). To keep computation tractable, we cap the Vietoris–Rips filtration at a distance threshold ϵ, which is swept as part of a robustness analysis.

We quantify the dissimilarity between two books (b, ℓ) and (b', ℓ') in homology dimension d as:

$$\delta^{(d)}_{(b,\ell),(b',\ell')} \;=\; \Delta\Big(X^{(d)}_{b,\ell},\, X^{(d)}_{b',\ell'}\Big), \tag{4}$$

where Δ is the Wasserstein distance between persistence diagrams, computed with **Persim** [8]. For the SL–SC baseline, we estimate within-book variability by drawing two independent subsamples of chapters from the same $X_{b,\ell}$ and computing $\delta^{(d)}$ between the resulting diagrams.

To validate that our findings are not artifacts of topological choices, we compute two non-topological metrics over the same 2×2 conditions.

For each language, we compute the pairwise cosine distance matrix among mean-pooled book embeddings. We then measure the agreement between

two languages by computing the Spearman correlation between the vectorized upper triangles of their respective distance matrices. A high correlation indicates that the inter-book similarity structure is preserved across languages independently of embedding geometry.

For each book embedding in language ℓ, we retrieve its nearest neighbor among book embeddings in language ℓ' and report Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR). We evaluate retrieval for both raw and centered embeddings, treating the difference as an ablation of the centering step.

4 Results

In this section, we present our findings on the preservation of structural and topological properties across languages in the GuoFeng Webnovel corpus. We first establish the alignment of the global embedding spaces using non-topological validation metrics, demonstrating the significant impact of language centering. We then present the results of our persistent homology analysis to evaluate whether higher-order geometric structures align with narrative content across language barriers.

To validate that our multilingual embeddings preserve inter-book relationships independently of topology, we analyzed the pairwise cosine distance matrices of mean-pooled book embeddings. The structural agreement across languages is exceptionally high and further improves after applying the language centering described in Section 3. By computing the Spearman correlation (ρ) between the flattened upper triangles of these distance matrices, we observed strong structural preservation, as detailed in Table 1.

Language Pair	Raw (ρ)	Centered (ρ)
EN vs. ZH	0.870	0.883
EN vs. DE	0.903	0.940
ZH vs. DE	0.854	0.869

Table 1: Spearman correlation of inter-book distance matrices across language pairs, comparing raw and centered embeddings.

This cross-lingual alignment is also dramatically reflected in the "same-book" retrieval performance. When retrieving the same book across languages using nearest-neighbor search, raw embeddings struggle with distant language pairs due to the global language offset. However, subtracting the per-language centroid effectively aligns the spaces. For EN $\rightarrow$ ZH, Top-1 retrieval accuracy surges from 0.382 to 0.974, with Top-5 reaching 1.000 and a Mean Reciprocal Rank (MRR) of approximately 0.983. For the ZH $\rightarrow$ DE and DE $\rightarrow$ ZH directions, Top-1 accuracy reaches a perfect 1.000 after centering. Notably, the EN $\leftrightarrow$ DE pair, being typologically closer, already achieved a Top-1 accuracy of 1.000 prior to centering.

To determine whether higher-order geometric features are preserved across translations, we evaluated our 2×2 factorial design using Vietoris–Rips persistent homology on the chapter point clouds ($X_{b,\ell}$). Following the methodology in Section 3, we report the Wasserstein distance (Δ) between persistence diagrams in the H_1 homology dimension using centered embeddings.

As hypothesized, the median distances between persistence diagrams follow a strict ordering that confirms topological alignment with content over language. The results for the four experimental conditions are presented in Table 2.

	Content	
Language	**Same (SC)**	**Different (DC)**
Same (SL)	3.36	5.40
Different (DL)	5.93	6.44

Table 2: Median Wasserstein distances (H_1) between chapter point cloud persistence diagrams across the 2×2 experimental conditions.

5 Conclusion

Crucially, we find that **DL–SC** $<$ **DL–DC** ($5.93 < 6.44$). This indicates that, after accounting for language offsets, the topological representation of the same book across different languages is significantly closer than the representations of different books across those same languages. The overall ordering (**SL–SC** $<$ **SL–DC** $<$ **DL–SC** $<$ **DL–DC**) demonstrates that while language still exerts a measurable influence on the topological shape of the point clouds (**SL–SC** $<$ **DL–SC**), the underlying narrative content preserves a distinct geometric signature that persists across translations. However, these results do not indicate that we can distinguish content based solely on topological similarity, since **SL–DC < DL–SC**. We are hopeful that further experimentation may reveal ways of associating topological structure more strongly with content.

References

[1] Ulrich Bauer. 2021. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423.

[2] Frédéric Chazal and Bertrand Michel. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963.

[3] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

[4] Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[5] J.R. Munkres. 1975. *Topology: a first course*. Prentice-Hall, New Jersey.

[6] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

[8] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. 2018. Ripser.py: A lean persistent homology library for python. In *The Journal of Open Source Software*.

[9] Longyue Wang, Siyou Liu, Minghao Wu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Liting Zhou, Yan Gu, Weiyu Chen, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the wmt 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*.

[10] Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, and 1 others. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67.