BEYOND WORDS: A TOPOLOGICAL EXPLORATION OF COHERENCE IN TEXT DOCUMENTS

Rishi Singhal*, **Samyak Jain***, **Sriram Krishna***, **Yaman Kumar Singla & Rajiv Ratn Shah**
Indraprastha Institute of Information Technology - Delhi

ABSTRACT

Coherence serves as a pivotal metric in evaluating the quality of a text. It quantifies how well the sentences within the text are connected and how well the text is structured and organized. It plays a vital role in various downstream Natural Language Processing tasks such as text summarization, question answering and machine translation among others. In this work, we explore the use of topological data analysis (TDA) techniques on attention graphs of text documents to model coherence. TDA techniques are known to capture structural information and patterns in data, making it suitable for modeling the *structure* and *flow* of a document, i.e. coherence. We validate our approach with experiments on the GCDC dataset, achieving state-of-the-art results with a simple MLP (code available publicly).

1 INTRODUCTION & RELATED WORK

Coherence is an essential feature of any well-organized text that explains the relationship between textual segments and shapes its flow. It is a measure of text quality, comprehensibility and logical consistency. Coherence assessment has significant implications in various tasks like text summarization (Parveen et al., 2016), machine translation (Mohiuddin et al., 2021) and language generation (Kiddon et al., 2016). Yet, modeling it is challenging due to its abstract nature, the ambiguities of language, and its subjective interpretation. Over time, various formal coherence models such as Centering Theory (Grosz et al., 1995) and Rhetorical Structure Theory (Mann & Thompson, 1987; Lin et al., 2011) have emerged, integrating concepts of syntactic features, co-references and lexical cohesion (see A.1). Classical methods however often overlook semantic aspects in text, leading to limitations. With the advent of deep learning, neural models like RNNs and LSTMs (Li & Hovy, 2014; Farag et al., 2020) which utilize the embedded semantic information, have advanced. More recently, Transformer based architectures have been used for coherence modeling and achieve state-of-the-art performance (Abhishek et al., 2022). Existing approaches in coherence modeling prioritize semantic accuracy but overlook text structure, potentially leading to incoherence even in semantically correct passages lacking cohesive flow. To address this gap, we take an integrated approach to utilizing Topological Data Analysis (TDA) techniques in conjunction with semantic information to model coherence. TDA is typically used to analyse sparse high-dimensional and noisy data that contain relevant low-dimensional features and patterns. Recent works utilize TDA based techniques for various domains like topic modeling (Byrne et al., 2022), artificial text detection (Kushnareva et al., 2021) and various text classification tasks (Doshi & Zadrozny, 2018; Elyasi & Moghadam, 2019). We make use of TDA methods on the attention maps extracted from a transformer to extract nuanced features that capture intricate structural and surface patterns within the text data. Built on contextualized embeddings from pre-trained language models, these features carry rich linguistic information and semantic properties. We show that simple neural networks trained on the TDA features computed from the attention maps outperform other neural baselines and achieve state-of-the-art performance in coherence modeling of text.

2 METHODOLOGY AND EXPERIMENTS

We extract attention matrices from a pre-trained language model (RoBERTa (Liu et al., 2019)) and transform them to a weighted graph where vertices represent tokens and edges represent the attention

*Equal Contribution. rishi19194@iiitd.ac.in, samyak19098@iiitd.ac.in

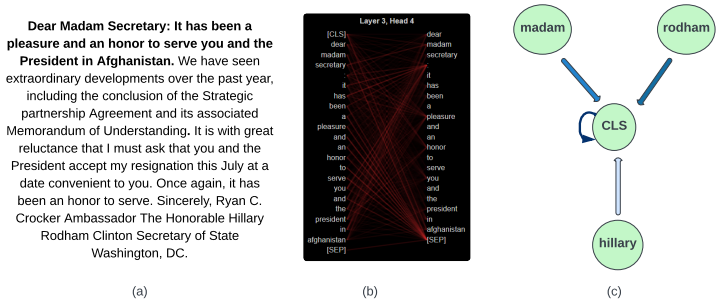

Figure 1: (a) Sample text from the GCDC dataset, (b) Attention map of a sentence in the document, (c) Thresholded attention graph of the document (darker color implies higher attention weight)

weights. We adopt the TDA-based modeling from Kushnareva et al. (2021) to generate features from the attention graphs: (i) **Topological features**: Standard directed graph properties like number of edges, cycles, connected components etc., which capture the global geometric information from the text's spatial representations; (ii) **Barcode features**: Descriptive characteristics of each *barcode* - a representation of the graph's persistent homology (Cherniavskii et al., 2022). These characterize the graph's *persistent* topological properties; (iii) **Distance to patterns**: Features based on distances of the text's attention maps to distinct attention patterns known to carry linguistic information (Clark et al., 2019). Detailed explanation of the three types of topological features are outlined in A.2.

These features are used to train a simple two-layer neural network for the downstream task of three-way classification for coherence assessment. The model is trained and tested on the Grammarly Corpus of Discourse Coherence (GCDC) dataset (Lai & Tetreault, 2018), a real world dataset, wherein each document is labelled as "low", "medium" or "high" based on the coherence of the text. The model architecture, dataset statistics and optimal parameters are outlined in detail in A.3 and A.4.

3 RESULTS

Model	Yahoo	Clinton	Enron	Yelp	Average
PARSEQ	54.9	60.2	53.2	54.4	55.7
Avg-XLNET-Doc	60.5	65.9	56.9	**59.0**	60.6
Fact-aware MTL	60.7	67.4	56.4	**59.0**	60.8
`TDA-MLP` (Ours)	**61.0**	**67.5**	**58.5**	57.5	**61.1**

Table 1: GCDC Dataset: Three-way classification results (accuracy in %)

Table 1 reports the accuracy scores on the GCDC dataset. Our model **`TDA-MLP`** is benchmarked against different baselines, including PARSEQ (Lai & Tetreault, 2018), Avg-XLNET-Doc (Jeon & Strube, 2020), and the state-of-the-art transformer model - Fact Aware MTL (Abhishek et al., 2022). **`TDA-MLP`** outperforms the different baselines in most domains, demonstrating the ability of the topological features to capture structural properties. The superior performance also validates the capacity of this method to mitigate noisy data patterns in the data, thus generating robust and highly stable discriminative features. We observe that our model also has a difficulty in classifying medium coherence samples, which we attribute to the inherent class imbalance in the data. (see A.5)

4 CONCLUSION

We introduce a topological perspective for the task of text coherence modeling. Our approach builds on the limitations of existing methods by utilizing a TDA-based features which capture structure and spatial information. Experimental results on the GCDC dataset show that a simple MLP trained using TDA-based features outperforms state-of-the-art transformers. Potential future directions can be aimed at investigating how TDA can complement existing methods in a few-shot setting or longer documents with distinct hierarchical structures and varying coherence definitions.

URM STATEMENT

The authors acknowledge that the key authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. Fact aware multi-task learning for text coherence modeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 340–353. Springer, 2022.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer (eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 141–148, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219858. URL https://aclanthology.org/P05-1018.

Ulrich Bauer. Ripser: efficient computation of vietoris–rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Ciarán Byrne, Danijela Horak, Karo Moilanen, and Amandla Mabona. Topic modeling with topological data analysis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11514–11533, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.792. URL https://aclanthology.org/2022.emnlp-main.792.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Acceptability judgements via examining the topology of attention maps. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 88–107, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.7. URL https://aclanthology.org/2022.findings-emnlp.7.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828.

Pratik Doshi and Wlodek Zadrozny. Movie genre detection using topological data analysis. In *International Conference on Statistical Language and Speech Processing*, 2018. URL https://api.semanticscholar.org/CorpusID:52955486.

Micha Elsner and Eugene Charniak. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pp. 41–44, USA, 2008. Association for Computational Linguistics.

Naiereh Elyasi and Mehdi Hosseini Moghadam. An introduction to a new text classification and visualization for natural language processing using topological data analysis, 2019.

Youmna Farag, Josef Valvoda, Helen Yannakoudakis, and Ted Briscoe. Analyzing neural discourse coherence models. *CoRR*, abs/2011.06306, 2020. URL https://arxiv.org/abs/2011.06306.

Vanessa Wei Feng and Graeme Hirst. Extending the entity-based coherence model with multiple ranks. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pp. 315–324, USA, 2012. Association for Computational Linguistics. ISBN 9781937284190.

Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45:61–75, 2007. URL https://api.semanticscholar.org/CorpusID:12894085.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995. URL https://aclanthology.org/J95-2003.

Stefan Horoi, Jessie Huang, Bastian Rieck, Guillaume Lajoie, Guy Wolf, and Smita Krishnaswamy. Exploring the geometry and topology of neural network loss landscapes. In *International Symposium on Intelligent Data Analysis*, pp. 171–184. Springer, 2022.

Sungho Jeon and Michael Strube. Incremental neural lexical coherence modeling. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6752–6758, 2020.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 329–339, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1032. URL https://aclanthology.org/D16-1032.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 635–649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.50. URL https://aclanthology.org/2021.emnlp-main.50.

Alice Lai and Joel Tetreault. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 214–223, 2018.

Jiwei Li and Eduard Hovy. A model of coherence based on distributed sentence representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2039–2048, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1218. URL https://aclanthology.org/D14-1218.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. Automatically evaluating text coherence using discourse relations. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 997–1006, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1100.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Annie Louis and Ani Nenkova. A coherence model based on syntactic patterns. In Jun'ichi Tsujii, James Henderson, and Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1157–1168, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1106.

William C Mann and Sandra A Thompson. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, 1987.

Mohsen Mesgar and Michael Strube. Graph-based coherence modeling for assessing readability. In Martha Palmer, Gemma Boleda, and Paolo Rosso (eds.), *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pp. 309–318, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-1036. URL https://aclanthology.org/S15-1036.

Tasnim Mohiuddin, Prathyusha Jwalapuram, Xiang Lin, and Shafiq R. Joty. Rethinking coherence modeling: Synthetic vs. downstream tasks. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 3528–3539. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EACL-MAIN.308. URL https://doi.org/10.18653/v1/2021.eacl-main.308.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991. URL https://aclanthology.org/J91-1002.

Daraksha Parveen, Mohsen Mesgar, and Michael Strube. Generating coherent summaries of scientific articles using coherence patterns. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 772–783, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1074. URL https://aclanthology.org/D16-1074.

Simon Zhang, Mengbai Xiao, and Hao Wang. Gpu-accelerated computation of vietoris-rips persistence barcodes, 2020.

A Appendix

A.1 Related Work: Classical Coherence Modeling

Preceding the emergence of deep learning, coherence modeling predominantly leveraged on entity-grid based models, proficiently encapsulating sentence structure, discourse entities, and grammatical transitions (Barzilay & Lapata, 2005). It breaks down a text into sequences of entities and evaluates coherence by modeling their relations to each other. Further advancements in this paradigm incorporate entity features (Elsner & Charniak, 2008), graph transformation (Mesgar & Strube, 2015), and refined ranking schemes (Feng & Hirst, 2012). However, these methods are unable to consider long transitions and model's inability to learn task specific features. Linguistic methods utilizing discourse relations, syntactic features and lexical cohesion exhibit effectiveness in coherence analysis. (Louis & Nenkova, 2012) show that sentences exhibit detectable structural patterns in a coherent text.(Morris & Hirst, 1991) use lexical chains to capture semantic context of structural elements of the text that contribute to coherence properties.

A.2 TDA Features Generation

Following the feature generation process outlined by (Kushnareva et al., 2021), we design three groups of features which are described in detail below.

Topological Features: These encompass various graph-based metrics derived from both directed and undirected graphs for multiple filtrations produced using a set of threshold values. A filtration for a graph for a particular threshold corresponds to a variant of the graph that only retains edges weighing more than the threshold value. Metrics measured includes edge count, strongly connected components, directed cycles, and average vertex degree. We also use the first two Betti numbers calculated from the undirected variant of the graphs. The final set of topological features are formed by concatenating these metrics for all thresholds.

Barcode Features: These set of features are extracted from barcodes of the first two persistent homology groups (H_0 and H_1) using `ripser++` Bauer (2021); Zhang et al. (2020). These barcodes in persistent homology represent topological features' lifespans across varying spatial scales for the data Ghrist (2007). The features we use encapsulate the following characteristics for the barcodes:

- Sum of length of bars
- Mean of length of bars
- Variance of length of bars
- Entropy of the barcode

- Time of birth of longest barcode (excluding infinite)
- Time of death of longest barcode (excluding infinite)
- Overall number of bars
- Total persistence
- Number of barcodes with time of death more than threshold value 0.75
- Number of barcodes with time of death more than threshold value 0.5
- Number of barcodes with time of death less than threshold value 0.25

Here, we also introduce a barcode feature called total persistence which is defined as the sum of squared lengths of barcodes, due to it being a powerful summary statistic, invariant towards rotation of embeddings and robust against perturbations of input filtration Horoi et al. (2022).

Distance to Patterns Features: Some patterns that appear in the attention graphs have been identified to carry linguistic information Clark et al. (2019). We calculate the distance to these patterns and use them as features from the graph. The distance between two graphs G_A and G_B with incidence matrices A and B is calculated as follows:

$$d(G_A, G_B) = \sqrt{\frac{\sum_{i,j}(a_{ij} - b_{ij})^2}{\sum_{i,j}(a_{ij}^2 + b_{ij}^2)}} \tag{1}$$

We get the attention graph G from the attention matrix and take the distance from the given graph G to these attention patterns G_i as the features, where G_i is:

- Attention to previous token: $G_i = E(i+1, i)$, $i = \overline{1, n-1}$
- Attention to next token: $G_i = E(i, i+1)$, $i = \overline{1, n-1}$
- Attention to [CLS]-token: $G_i = E(i, 1)$, $i = \overline{1, n}$
- Attention to [SEP]-token: $G_i = E(i, i_t)$, $i = \overline{1, n}$, $t = \overline{1, k}$ where i_1, i_2 ... i_k are indices of [SEP]-token.
- Attention to punctuation marks: G_i = E(i, i_t) , i = $\overline{1, n}$, t = $\overline{1, k}$ where i_1, i_2 ... i_k are indices of tokens corresponding to commas and periods.

A.3 Dataset Description

The Grammarly Corpus of Discourse Coherence (GCDC) comprises diverse texts, including emails and reviews, spanning domains such as Yahoo forum posts, Hillary Clinton's office emails, Yelp reviews, and Enron emails. Unlike synthetic low-coherence documents, GCDC features authentic texts with varying language proficiency. Expert annotations categorize documents into low, medium, and high coherence, transforming GCDC into a 3-way classification task. Dataset statistics are summarized in Table 2. We can see that the dataset is imbalanced i.e. - there are a low number of "Medium" coherence samples.

Domain	#Docs	Avg #Words	Avg #Sents	Low, Medium, High (%)
Yahoo	1200	162.1	7.5	46.6, 17.4, 37.0
Clinton	1200	189.0	6.6	28.2, 20.6, 51.2
Enron	1200	196.2	7.7	29.9, 19.4, 50.7
Yelp	1200	183.1	7.5	27.1, 21.8, 51.1

Table 2: GCDC dataset statistics

A.4 Model Description

Following the extraction of the TDA features for each document, we consolidate them into a single feature vector of size 9216 x 1 and pass it through a simple 2-layer MLP model as described in Figure 2

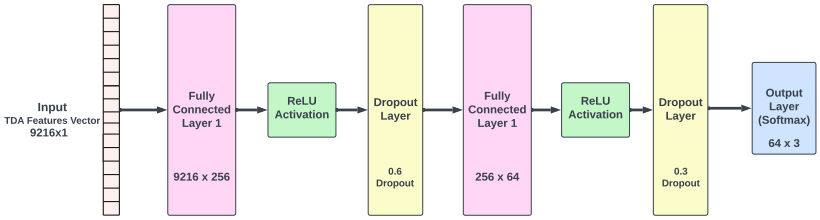

Figure 2: MLP Model Architecture

Also, this MLP model is trained using the Adam optimizer, employing Cross Entropy Loss as the objective function with key hyperparameters as 100 epochs, learning rate set of $1e-5$, weight decay of $5e-2$, and epsilon (used to control numerical stability in Adam optimizer) configured to $1e-8$. These optimal set of hyperparameters are obtained using hyperparameter tuning via Grid Search method. The results reported for our model are averaged over 10 runs to ensure stability.

A.5 Analysis of Results

Table 3 contains a detailed report of multiple performance metrics across each of the coherence classes in GCDC (Low, Medium, High). **TDA-MLP** performs very poorly on the Medium Coherence class, with an average recall of 2.75, indicating that most samples belonging to this class are misclassified as High/Low coherence. A promising direction for future research is to verify if other methods like Fact-aware MTL Abhishek et al. (2022) also perform poorly on the Medium Coherence class and investigate the reason.

Metrics	Yahoo	Clinton	Enron	Yelp	Average
Accuracy	76.8, 0.0, 75.3	54.9, 2.6, 92.7	70.9, 2.0, 87.5	36.0, 2.3, 88.8	59.6, 1.7, 86.1
Precision	63.0, 0.0, 59.0	50.0, 100.0, 72.0	70.0, 25.0, 58.0	51.0, 50.0, 59.0	58.5, 43.75, 62.0
Recall	77.0, 0.0, 75.0	55.0, 3.0, 93.0	71.0, 2.0, 88.0	36.0, 2.0, 89.0	59.7, 2.75, 86.25
F1-Score	69.0, 0.0, 66.0	52.0, 5.0, 81.0	70.0, 4.0, 70.0	42.0, 5.0, 71.0	58.25, 3.5, 72.0

Table 3: Performance Metrics for **TDA-MLP** on GCDC. Results are reported as a triple of <Low, Medium, High> coherence samples.

Our model **TDA-MLP** takes as input 3 kinds of features: 1) Topological Features 2) Barcode Features 3) Distance-to-Patterns features. (described in A.2). Table 4 contains the results for each of these features independently with the same architecture.

	Yahoo	Clinton	Enron	Yelp	Average
Topological	39.5	54.5	48.0	48.0	47.5
Barcode	37.5	52.5	45.5	50.0	46.37
Distance-to-Patterns	38.5	52.5	45.5	54.0	47.62
TDA-MLP	**61.0**	**67.5**	**58.5**	**57.5**	**61.1**

Table 4: Ablation studies for **TDA-MLP**

We observe that each of the three feature groups perform passably well on the GCDC, indicating that each group captures coherence differently. When put together, **TDA-MLP** shows strong performance on the GCDC dataset, despite using a simple MLP.