Topology in Neural Machine Translation: A Topological Study of Transformers through Attention

Yuwei (Johnny) Meng

3 Nov 2025

1 Introduction

In 2017, a group of researchers at Google proudly announced the architecture of the transformer neural model, which revolutionized the field of Natural Language Processing (NLP) (Vaswani et al., 2017). Before then, neural NLP models mainly relied on recurrent structures, such as RNNs and LSTMs, optionally with the attention mechanism to boost performance. The transformer architecture, however, abandoned the recurrent structure in RNNs and LSTMs and solely utilized attention for language modeling, which was a novel but successful approach. Since the invention of transformer, NLP researchers have been applying this architecture to various NLP tasks, one of which is Machine Translation (MT). MT is an NLP task that takes a sentence in a source language as input and outputs the translated sentence in a target language, and Neural Machine Translation (NMT) is a subfield of MT that specifically uses neural networks as the model for the translation. According to Vaswani et al. (2017), the transformer model achieved a BLEU score of 41.0 on the WMT14 English-to-French benchmark, establishing a new state-of-the-art performance.

Despite the prominent performance of transformer, similar to other neural network architectures, the specific reasons behind its success remain largely unknown, particularly in the context of NMT. One method to probe the interpretability of neural networks is through Topological Data Analysis (TDA), where topological features that are intrinsic to the data and model are extracted and explained. This method, nevertheless, is also underexplored in NMT. Therefore, in this research project, I propose to apply TDA to explain the power of transformer on the task of NMT. Since the attention mechanism is the core of transformer, topology-related techniques will be applied to analyze the attention maps generated by transformer models during translation. Considering my

knowledge of English and French and the abundance of such evaluation datasets, I will focus on the English-to-French translation task. The specific research question is as follows:

> *Do English and French sentences create similar topological structures in the attention maps generated by transformer models during translation? If so, does similarity in the topological structures of attention maps correlate with translation quality?*

2 Related Work

There are past studies that focused on using algebraic topology to analyze neural networks. For example, the paper by Bianchini and Scarselli (2014) is a pioneer study that leveraged topological tools to compare the expressivity of shallow and deep neural networks. They discovered that for deep neural networks, the sum of the Betti numbers, as a metric that measures the topological complexity that the network can express, can grow exponentially with the number of hidden units. Later, Guss and Salakhutdinov (2018) extended this work by empirically applying Betti numbers to measure the topological complexity of real-world datasets and characterize the expressivity of fully-connected neural networks. These studies laid a solid foundation for using topological tools to study neural networks, but the generalization of such techniques to more complex neural structures still remains limited.

In addition to studying neural networks using topology, there are also attempts to apply topological techniques on language modeling. Fitz (2022) introduces the notion of a *word manifold*, which turns n-gram models on raw texts of various languages into simplicial complexes, allowing for topological analysis. This study differs from my proposed study in that the method is applied to raw texts with no neural models associated with it. More recently, Draganov and Skiena (2024) makes an effort to study word embeddings generated by large language models by considering the d-dimensional space that these embeddings are located in as a topological space. Then, they applied persistent homology to extract topological patterns from the embedding spaces formed by 81 languages. Their study suggested statistically significant results that word embeddings carry meaningful linguistic information, but there was no analysis of the underlying neural models.

Perhaps the most closely related study to my proposed topic is the one by Haim Meirom and Bobrowski (2022). In this study, they also looked at embeddings as Draganov and Skiena (2024) did, but they argue that certain semantics are inherent to the real world and are not language dependent. For example, *dog* and *cat* are both common pets so they often appear in the same context regardless of the language. Under this assumption, they claimed that the embedding spaces

of different languages should be isomorphic to each other at the sentence level, and their results supported this. An interesting question therefore arose from this study: since the embedding spaces of different languages at the sentence level are isomorphic, and an NMT system transforms a sentence from the source language to the target language, how does the NMT system preserve such isomorphism during translation? This question is thus the main motivation of my proposed research.

Now, is looking at attention feasible? Previous studies said yes. Ravishankar et al. (2021) studied fully using the attention of multilingual BERT to decode syntactic dependency trees of 18 languages, including English and French, and their results showed that solely using attention can achieve competitive accuracy in dependency parsing, suggesting that attention does encode meaningful syntactic information, which could be helpful in translation as well. Furthermore, Kushnareva et al. (2021) studied the attention mechanism with topology. In the study, they first built weighted graphs from attention maps by treating tokens as nodes and attention weights as edges, followed by applying persistent homology on the graph to construct a filtration. Their topic was on detecting artificially generated texts which is different from mine, but this process is inspiring that I will adopt in my study.

To conclude this section, Uchendu and Le (2025) in their paper of a survey on using TDA to approach NLP problems stated that:

> *One glaring application is on multi-lingual tasks... Due to the benefits of TDA which include performing robustly on heterogeneous, imbalanced, and noisy data, its application on multi-lingual tasks is necessary.*

This statement further motivates my proposed topic.

3 Methodology

The proposed methodology is as follows:

1. Choose an NMT model. The NLLB (No Language Left Behind) model developed by Meta is a good candidate (Team et al., 2022). This model offers translation between 200 languages, including English and French, and is open-sourced. The NLLB models are available on Hugging Face ranging from 600M to 54B parameters, allowing me to experiment with different model sizes considering possible computational constraints.
2. Pick an evaluation dataset that aligns English and French sentences. The evaluation sets

curated by the *Workshop on Machine Translation* (WMT) are good resources to use and are available on Hugging Face, among which WMT14 is a solid choice which was also selected by Vaswani et al. (2017).

3. Choose > 1000 sentences from the datasets. Run the NMT model on these sentences to obtain the self-attention maps for both the English encoder and the French decoder.
4. For each attention map, build a weighted graph by treating tokens as nodes and attention weights as edges, following Kushnareva et al. (2021). Run persistent homology on the weighted graph to extract topological features ($\beta_0 =$ connected components and $\beta_1 =$ loops).
5. Align the self-attention topological features of the encoder and decoder based on sentence pairs. Compute the similarity between the features and find correlations between similarity scores and translation quality, measured by BLEU scores.
6. Optionally track changes in topological features throughout the transformer layers.
7. If time permits, experiment with other language pairs such as English-to-Chinese.

The proposed research topic has its significance because of the following reasons:

- If we do find that topological features are similar between English and French sentences, then it suggests that the transformer models are capable of preserving topological structures during translation, which helps explain their success in NMT.
- If similarity in topological features correlates with translation quality (i.e. BLEU scores), then we can possibly develop topology-based metrics to evaluate translation quality in the future.
- However, if the features differ, it still suggests that more work on other perspectives needs to be done to fully interpret an NMT system that uses transformer.

I hope this research can shed some light on the interpretability of transformer models in NMT through topological analysis. The findings may also inspire future research that applies topology to other NLP tasks involving transformer models.

References

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, 2014.

Ondřej Draganov and Steven Skiena. The shape of word embeddings: Quantifying non-isometry with topological data analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12080–12099, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Stephen Fitz. The shape of words - topological structure in natural language data. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 116–123. PMLR, 25 Feb–22 Jul 2022.

William H. Guss and Ruslan Salakhutdinov. On characterizing the capacity of neural networks using algebraic topology. *CoRR*, abs/1802.04443, 2018.

Shaked Haim Meirom and Omer Bobrowski. Unsupervised geometric and topological approaches for cross-lingual sentence representation and comparison. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 173–183, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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*, pages 635–649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online, April 2021. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov,

Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

Adaku Uchendu and Thai Le. Unveiling topological structures from language: A comprehensive survey of topological data analysis applications in nlp, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.