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

Yuwei (Johnny) Meng

23 Dec 2025

Abstract

1. Introduction

2. Background

2.1. Algebraic Topology

Topology is a branch of mathematics that characterizes shapes, spaces, and sets by their connectivity (Guss & Salakhutdinov, 2018). Algebraic topology, more sophisticatedly, is a subfield of topology that attributes algebraic properties such as groups and chains to topological spaces in order to make explanations and interpretations more expressive. Formally, let X be a compact metric space. We can define a p-simplex to be a collection of points $\{x_0, \cdots, x_n\} \subseteq X$ in p-dimension. Depending on the value of p, these simplices bear different names:

- $p = 0$: point
- $p = 1$: line
- $p = 2$: triangle
- $p = 3$: tetrahedron
- $\cdots$

Now consider a collection of such p-simplices, called $\mathcal{K}$. Then $\mathcal{K}$ is called a *simplicial* complex if it satisfies these conditions:

1. If $\sigma \in \mathcal{K}$ and τ is a face of σ, then $\tau \in \mathcal{K}$;
2. If $\sigma_1, \sigma_2 \in \mathcal{K}$, then $\sigma_1 \cap \sigma_2 = \emptyset$ or $\sigma_1 \cap \sigma_2 \in \mathcal{K}$.

Given a simplicial complex $\mathcal{K}$ in the compact metric space X, one method that is frequently used to study $\mathcal{K}$ is homology. The core idea of homology is to construct chains, cycles, and boundaries from the simplices in $\mathcal{K}$ and analyze their relationships. Given a dimension n, the nth homology group

of the compact metric space X is defined as $H_n(X) = \mathbb{Z}^{\beta_n}$, where β_n is called the *nth Betti number*. For $n \geq 1$, the nth Betti number β_n measures the number of n-dimensional holes in the simplicial complex $\mathcal{K}$, while β_0 measures the number of connected components in $\mathcal{K}$. For example, a torus is a 3-dimensional object with 1 connected component, 2 1-dimensional holes, and 1 2-dimensional void. Therefore, the homology groups of the torus are $H_0(X) = \mathbb{Z}^1$, $H_1(X) = \mathbb{Z}^2$, and $H_2(X) = \mathbb{Z}^1$.

2.2. Persistent Homology

Realistically, given a collection of n-dimensional points in $\mathbb{R}^n$, we would like to extract meaningful topological information that characterizes these points. Persistent homology is thus one method that computes topological characteristics of this collection of points. Given a collection of points P, we first construct a simplicial complex $\mathcal{K}$ known as the Vietoris-Rips (VR) complex. The vertices in $\mathcal{K}$ are just the points in P. To build the edges that connect the points, we consider an increasing sequence of radii. For each radius r in the sequence, we superimpose a circle of radius r on each point in P. If for some radius r_1 the circle at point x_1 starts to cover another point x_2, then we connect x_1 and x_2 by an edge at r_1.

Notice that as the radius r increases, more and more edges would be connected, leading to emergence and disappearance of topological features. We thus can characterize these topological features by their emergence time and disappearance time, or birth time and death time using standard topology terminology. For instance, recall our previous example concerning points x_1 and x_2. If a 1-dimensional hole α emerges because of the addition of the edge between them at r_1, then we would say that α has a birth time of r_1. Similarly, if at some later radius $r_2 > r_1$ that α disappears because of adding another edge in the simplicial complex, then we would denote r_2 as the death time of α.

Figure 1 below shows a visualization of computing persistent homology using the method described above on a set of points in $\mathbb{R}^2$. The left figure shows the sequence of radii increasing from 0 to 6.15, along which edges are added to the simplicial complex. Note that at $r = 5.6$, the addition of the edge between p_1 and p_3 make the simplicial complex into a quadrilateral, which results in a 1-dimensional hole. Subsequently, at $r = 6.15$, the quadrilateral is destroyed by the edge between p_1 and p_4, causing the 1-dimensional hole to disappear. The simplicial complex constructed by increasing the radius r is called a *filtration*. On the other hand, the right figure shows the persistence diagram of this filtration. Note that the 1-dimensional feature described previously corresponds to the orange point at $(5.6, 6.15)$ in the persistence diagram.

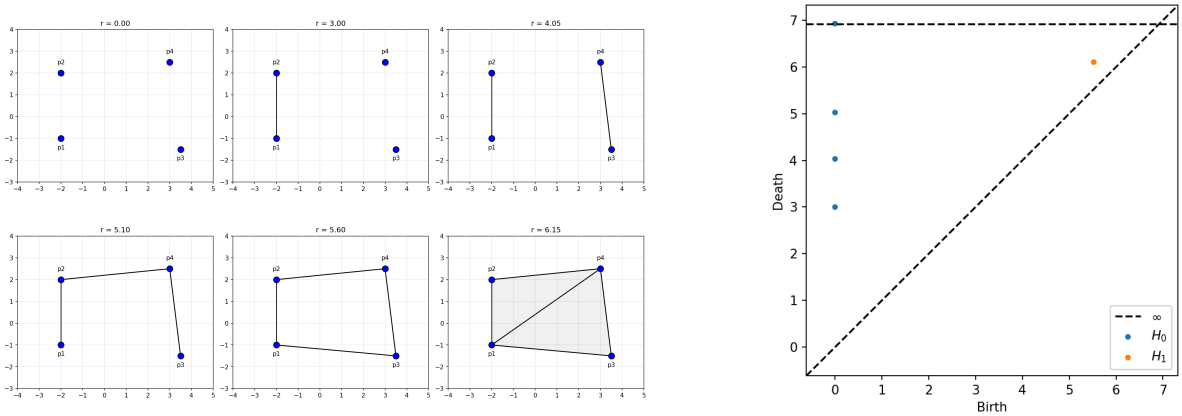

Figure 1: Visualization of Vietoris-Rips filtration on a set of points in $\mathbb{R}^2$.

Now given two sets of points, we can compute separately their persistence diagrams using persistent homology, but we need a metric to compare the two persistence diagrams. *Wasserstein distance* is the most common metric for this task. Given persistence diagrams D_1 and D_2, Wasserstein distance finds the best bijection between points in D_1 and D_2 and computes the sum of distances between matched points. Formally, let p be a fixed dimension. The p-Wasserstein distance between D_1 and D_2 is defined as:

$$W_p(D_1, D_2) = \inf_{\phi: D_1 \to D_2} \left(\sum_{x \in D_1} \|x - \phi(x)\|^p \right)^{1/p}.$$

The Wasserstein distance between D_1 and D_2 is thus:

$$W(D_1, D_2) = \sum_{n=0}^{\infty} W_n(D_1, D_2).$$

3. Related Work

There are numerous past studies that focused on using algebraic topology to analyze neural networks. For example, the paper by Bianchini & 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 & 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 & 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 Meirom & Bobrowski (2022). In this study, they also looked at embeddings as Draganov & 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 & 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.*

Hence, the current study of applying TDA on NMT systems is motivated.

4. Methodology

4.1. Model

This study revolves around studying the attention mechanisms of NMT systems. Therefore, an NMT system must be selected so that we can extract the attention maps to analyze. In this study, the NLLB (No Language Left Behind) model developed by Meta was chosen (Costa-jussà et al., 2022). This model offers translation between 200 languages, including English, French and Chinese, and is open-sourced. The NLLB models are available on Hugging Face ranging from 600M to 54B parameters. The distilled NLLB with 1.3B parameters was selected for its balance between performance and computational cost. This model has 24 encoder layers and 24 decoder layers, each layer containing 16 attention heads.

4.2. Datasets

The selected model must be run on a corpus to generate the attention maps. I picked the evaluation datasets curated by the *Workshop on Machine Translation (WMT)*. WMT is a well-known NMT benchmark that hosts annual MT competitions, each year with a different combination of source and target languages. In this study, the 2014 WMT (WMT14) benchmark is chosen for the French-English analysis and the 2017 WMT (WMT17) benchmark is chosen for the Chinese-English analysis due to their popularity. Vaswani et al. (2017) also used WMT14 to evaluate their transformer model.

The WMT14 French-English benchmark contains 3,000 validation sentence pairs, while the WMT17 Chinese-English benchmark contains only 2,000 validation sentence pairs. To ensure a fair comparison between languages, I selected only the first 2,000 sentence pairs from the WMT14 French-English validation dataset, and all 2,000 sentence pairs from the WMT17 Chinese-English validation dataset to conduct the experiment.

4.3. Experiment

We would like to analyze whether the attention maps generated by the NMT model are different for each language. Therefore, for each French-English sentence pair in the datasets, the model is run on the English sentence to generate the French translation, as well as on the French sentence to generate the English translation. This way, the encoder of NLLB would process the same sentence twice but in two different languages, and so we can extract the encoder attention map for both the English and French sentences. The experiment is repeated for the Chinese-English sentence pairs. Upon generating the translations, only the attention maps in the last layer of the encoder are extracted for analysis, since the last layer is expected to contain the most refined attention information. The attention weights across the 16 heads are mean-aggregated to form a single attention map for each sentence.

After the attention maps for all the sentence pairs are generated and extracted, we apply topological methods to analyze these attention maps. For each attention map, we build a weighted graph by treating the words in the sentence as nodes and adapting the attention weights as edges, following Kushnareva et al. (2021). The weight of each edge has the value of $1 - \alpha$, where α is the attention weight between two words. This way, a higher attention weight would correspond to a smaller edge weight, resembling a shorter distance between two nodes. This method allows a more intuitive interpretation of the Vietoris-Rips complex constructed later.

Upon building the weighted graph for each attention map, we run the filtration process described in *Section 2.2* to compute the persistence diagram using persistent homology. Consequently, for each sentence pair, this step results in two persistence diagrams, one for the English sentence and one for the sentence in the other language (French or Chinese). In this study, only zeroth-order and first-order topological features are considered for two reasons. First, the French-English dataset has mean sentence lengths of 17.9 tokens for English and 19.4 tokens for French, while the means for the Chinese-English dataset are 26.0 for English and 43.9 for Chinese. These numbers indicate that the VR complexes for the sentence pairs are relatively small, making higher-order topological features less likely to appear. Second, zeroth-order and first-order features are easier to interpret in the context of attention maps than higher-order features, which may not have clear meanings in this setting.

The corresponding computed persistence diagrams for each sentence pair permit the calculation of the Wasserstein distance between the two diagrams. From the Wasserstein distance, we learn how NMT systems attend to different languages differently at a topological level. In this context, a smaller Wasserstein distance indicates that the attention maps for the two languages are topologically similar,

which shows that the NMT system processes the two languages in a similar manner. Conversely, a larger Wasserstein distance indicates that the attention maps are topologically different, suggesting that the NMT system treats the two languages differently.

Lastly, we would like to see if topological differences in attention maps can be an indication of translation quality. Therefore, given the translation generated in a previous step of the experiment, we compute the BLEU score by comparing the translation to the reference sentence in the dataset. For each language direction, we aggregate the BLEU score information and conduct correlation analysis with the Wasserstein distances computed previously. For this step, we hypothesize a strong negative correlation between BLEU scores and Wasserstein distances, as it is intuitive that better translations should correspond to more similar attention maps between the two languages.

5. Results & Discussion

This section presents the topological findings from the experiment described above. *Section 5.1* shows the analysis of attention maps using Wasserstein distance, delving into whether transformer models attend to sentences of different languages differently. Next, *Section 5.2* presents some insights into the translation quality of the NMT system used in this study. Then, *Section 5.3* combines the results from *Section 5.1* and *Section 5.2* and conducts correlation analysis to see if topological differences and translation quality are correlated. Lastly, *Section 5.4* concludes this section by suggesting some possible explanations for the observed results, further enhancing the interpretability of the findings.

5.1. Topology

Before analyzing the topological features in the attention maps, let's examine a typical attention map. The left plot of *Figure 2* below shows the averaged self-attention in the last layer of the encoder for an English sentence. From the plot, note that the NLLB model creates a language tag at the beginning of the sentence, as well as an end-of-sentence (EOS) tag at the end of the sentence. Further notice that the EOS column of the plot has a very light color, meaning that every token in the sentence seems to attend strongly to the EOS token. This phenomenon suggests that the model considers the EOS token to be very important, perhaps using it as a scratchpad to store information about the entire sentence. The pattern is also present for the language tag, but not as strong as EOS. However, since we are analyzing how the NMT model understands the sentence, not byproducts of the translation process, these special tags are removed and the remaining attention weights are renormalized to 1 for the topological analysis. The middle plot of *Figure 2* shows the attention map after filtering and renormalization. Additionally, the right plot of *Figure 2* shows the distance matrix of the sentence,

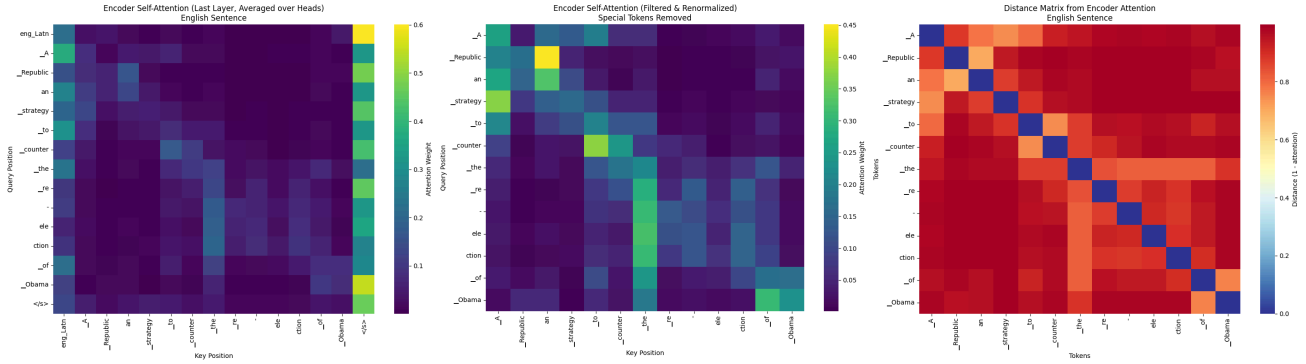

Figure 2: Attention maps and distance matrix for the English sentence “A Republican strategy to counter the re-election of Obama”.

where each distance has the value of $1 - \alpha$, with α being the attention weight between two words, and the distance of every word to itself is set to 0.

With the distance matrix, the persistence diagram of this attention map can be computed. *Figure 3* below shows the persistence diagrams of the same sentence presented in *Figure 2* in English and French. Note that both persistence diagrams show various zeroth-order topological features, which is expected because different parts of the sentence attend differently to each other. However, first-order topological features seem to be rare and ephemeral in both diagrams, indicating that there are not many loops in the attention maps, and loops tend to be short-lived. These patterns are consistent across most sentences, both in the French-English and Chinese-English datasets.

Table 1 below shows the summary statistics for the metrics computed for both the French-English and Chinese-English datasets. The table presents the minimum, maximum, mean, and median values for Wasserstein distances, token counts, and topological features across the 2,000 sentence pairs for each language pair. From the table, note that the majority of topological differences between languages seem to stem from zeroth-order topological features, as we can see that first-order topological features only contribute to smaller than 1.0 Wasserstein distance on average for both datasets. This discovery aligns with the previous observation that first-order topological features are rare in attention maps. Furthermore, the French sentences in our dataset are generally longer than their English counterparts, but the Chinese sentences have about the same number of tokens as their English counterparts. Nevertheless, the model still generates more H_1 features for Chinese sentences than for English sentences on average, indicating that the model attends to Chinese sentences in a more complex manner.

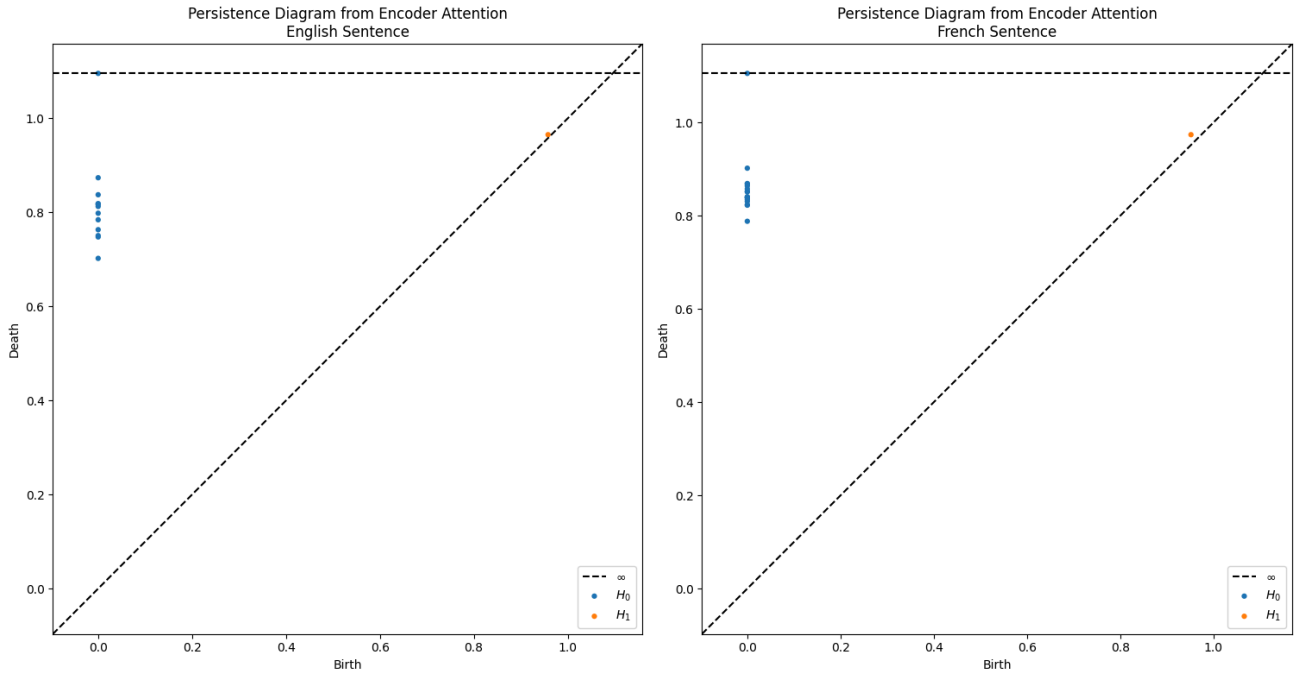

Figure 3: Persistence diagrams for the English sentence "A Republican strategy to counter the re-election of Obama" (left) and its French translation "Une stratégie républicaine pour contrer la réélection d'Obama" (right).

5.2. Translation Quality

Now we examine the translation quality of the NLLB model on the datasets with BLEU scores. The BLEU metric is a widely-used measurement of translation quality that compares the generated translation to a reference sentence. The BLEU score is comprised of a brevity penalty factor that penalizes the translation from being too short, as well as a geometric mean of modified n-gram precisions that measures how many n-grams in the translation appear in the reference sentence. The final BLEU score ranges from 0 to 100, with higher scores indicating better translation quality (Papineni et al., 2002).

Figure 4 below shows the distributions of BLEU scores for translations in the French-English and Chinese-English datasets. From the plots, we note that the Chinese-English language pair achieves lower BLEU scores on average than the French-English, by approximately 10 points. In particular, perfect translation (BLEU $= 100$) is not uncommon between French and English but rarely seen in Chinese-English translations. This phenomenon is expected because French and English have the same typological roots, while Chinese belongs to a completely different language family. Therefore, it is generally more difficult to translate between Chinese and English than between French and English. A more specific analysis of translation errors is presented in *Section 5.4*.

Metric	Min	Max	Mean	Median
French-English				
Wasserstein Distance (Total)	0.0	40.9	5.0	4.1
Wasserstein Distance (H_0)	0.0	40.3	5.0	4.1
Wasserstein Distance (H_1)	0.0	0.5	0.1	0.0
Token Count (H_0 Features) (English)	1	117	24.7	22
Token Count (H_0 Features) (French)	2	177	31.7	28
H_1 Features (English)	0	51	4.6	3
H_1 Features (French)	0	78	5.7	4
Chinese-English				
Wasserstein Distance (Total)	0.2	26.6	4.4	3.5
Wasserstein Distance (H_0)	0.2	26.2	4.2	3.3
Wasserstein Distance (H_1)	0.0	0.7	0.1	0.1
Token Count (H_0 Features) (English)	2	109	36.0	35
Token Count (H_0 Features) (Chinese)	4	111	36.2	35
H_1 Features (English)	0	41	8.4	8
H_1 Features (Chinese)	0	67	12.6	11

Table 1: Summary statistics for topological metrics computed on the French-English and Chinese-English datasets (2,000 sentence pairs each).

5.3. Correlation Analysis

After computing the Wasserstein distances and BLEU scores for all 2,000 sentence pairs in the datasets, we conduct a correlation analysis to examine whether Wasserstein distance is an indication of translation quality. *Figure 5* below shows the scatter plots of Wasserstein distances and BLEU scores for the datasets. We notice that in all 6 plots, the Pearson correlation coefficients are all negative, but the magnitudes are all smaller than 0.1. All the coefficients are statistically significant at the significance level of 0.05, with p-values less than 0.05. These results suggest that, although the correlations are weak, there is significant evidence that there are negative correlations between Wasserstein distances and BLEU scores in all language directions. This finding partially supports our earlier hypothesis that there is a negative correlation between topological differences and translation quality, as better translations should correspond to more similar attention maps. However, the correlation is not as strong as expected.

Now, as shown in *Table 1* above, token count directly reflect the number of zeroth-order topological features in the sentence, which can possibly confound the correlation analysis of Wasserstein

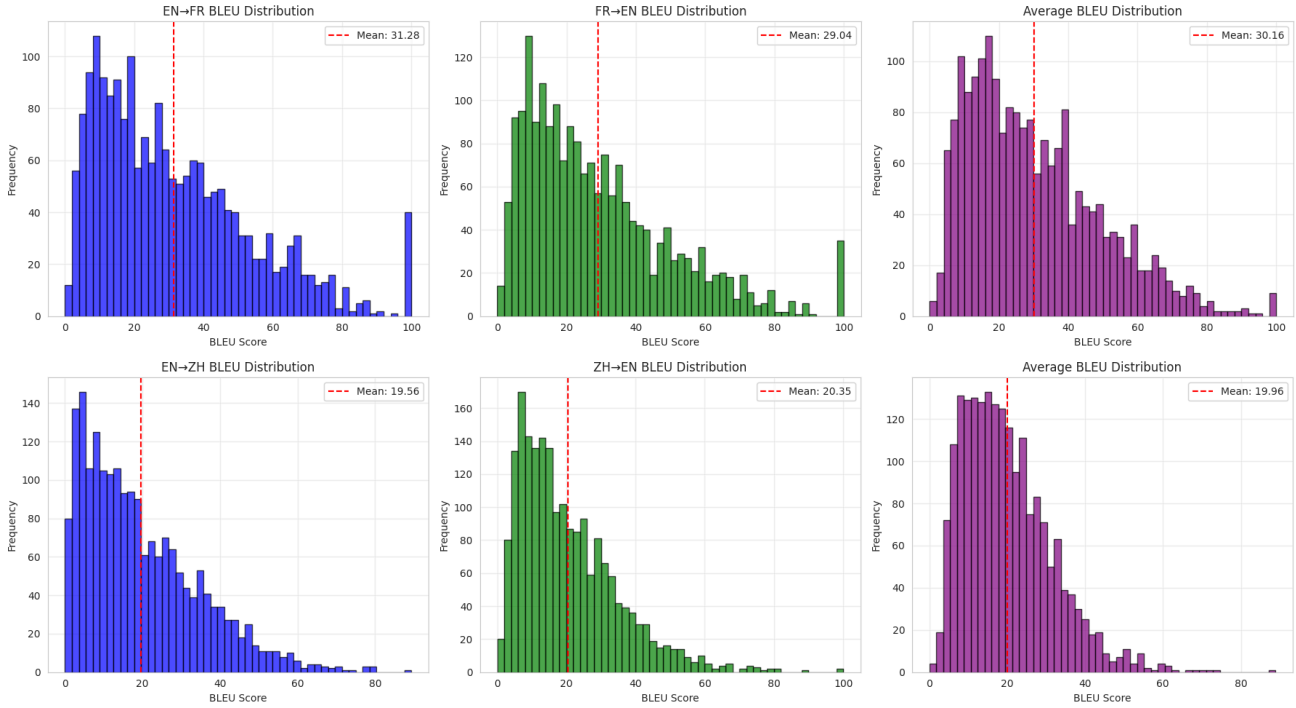

Figure 4: Distributions of BLEU scores for translations in the French-English (top) and Chinese-English (bottom) datasets.

distances and BLEU scores. This factor is also intuitively worrisome because longer sentences

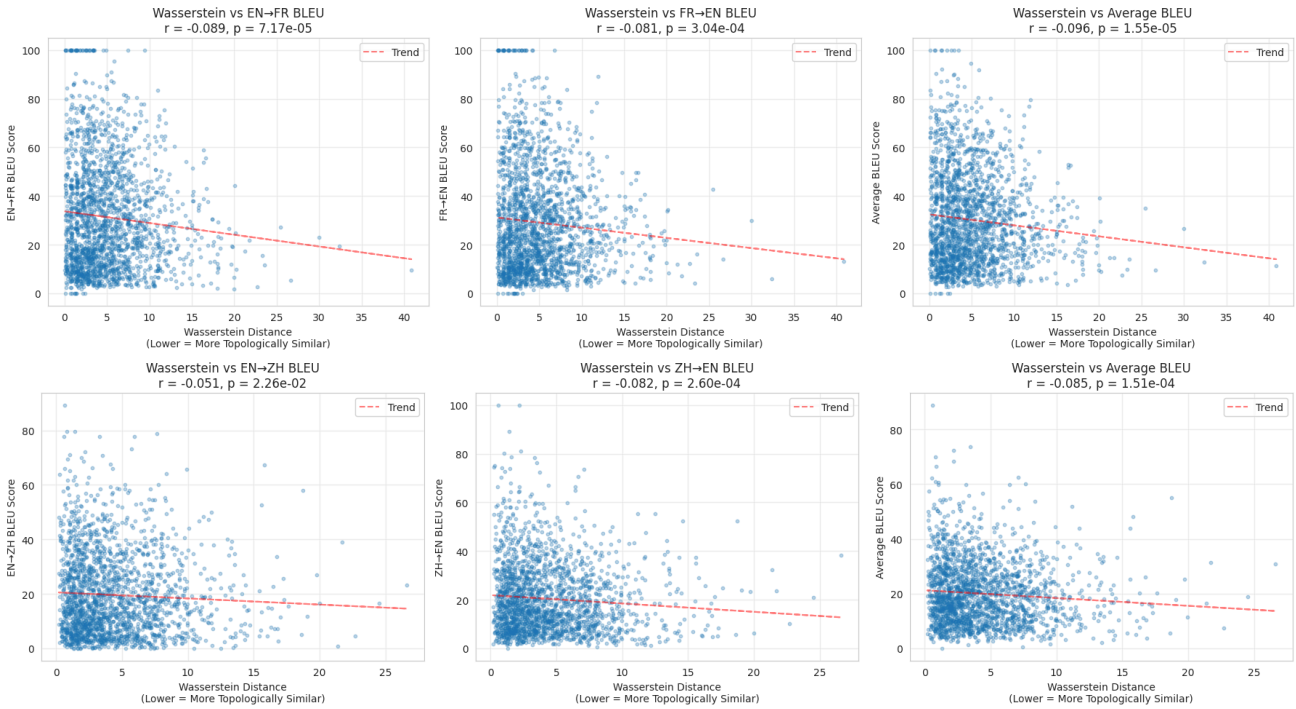

Figure 5: Scatter plots showing the relationship between Wasserstein distances and BLEU scores for the French-English (top) and Chinese-English (bottom) datasets.

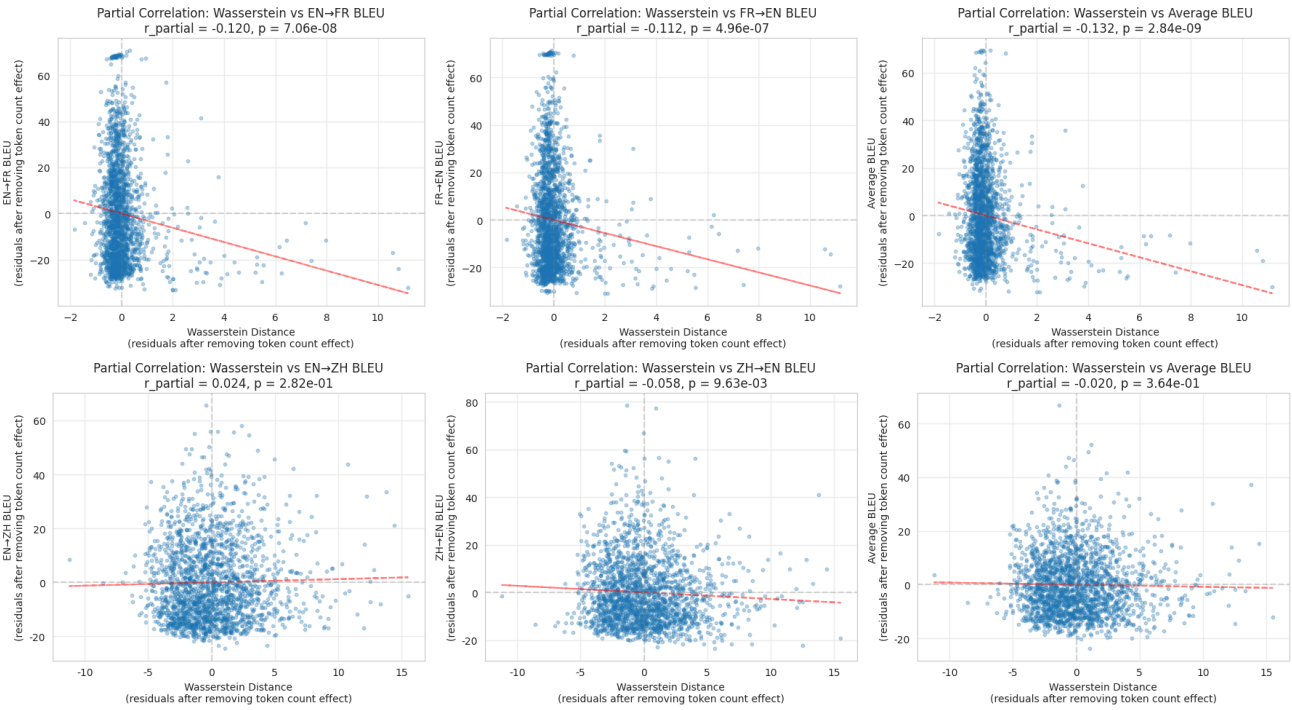

Figure 6: Scatter plots showing the relationship between Wasserstein distances and BLEU scores after controlling for token counts for the French-English (top) and Chinese-English (bottom) datasets.

are generally more difficult to translate correctly, hence lowering translation quality. Therefore, a correlation analysis that controls for the effect of token count is necessary. *Figure 6* below shows the scatter plots for this purpose.

Figure 6 shows the scatter plots for the partial correlation between Wasserstein distances and BLEU scores. Given the two variables of interest for correlation analysis and one or more possibly confounding variables, partial correlation first fits two linear regression models that use the confounding variables to predict each variable of interest separately. With the linear regression models comes the residuals of the two variables of interest that the confounding variables cannot explain. Then, we carry out the correlation analysis on the two sets of residuals, attempting to find correlation between the two variables of interest in the parts that are not affected by the confounding variables. In the context of this paper, the two variables of interest are the Wasserstein distances and the BLEU scores, while the only confounding variable is the token count.

From *Figure 6*, we note that the partial correlations for the French-English pair are higher than the correlations from *Figure 5* in both language directions, and the partial correlation for the averaged case is the highest. These results are also strongly statistically significant with very small p-values. This suggests that, even though French and English sentences have very different token counts as indicated in *Table 1*, this does not seem to confound the correlation between topological differences in

the attention maps and translation quality. In fact, controlling for token count renders the correlation stronger, which shows that token count is masking the true correlation between Wasserstein distances and BLEU scores. Therefore, we can conclude that preserving topological features in attention maps independently contributes to translation quality between French and English.

On the other hand, the partial correlations for the Chinese-English pair become weaker as shown in *Figure 6*, with statistically insignificant p-values in the English-Chinese direction and the averaged case. This shows that token count, in the Chinese-English case, is strongly confounding with the correlation between Wasserstein distances and BLEU scores. In other words, the difficulty in translating longer sentences stands out more in this case, possibly overwriting the topological differences in the attention maps that the model generates. We will try to provide reasons for this outcome in *Section 5.4* below.

5.4. Error Analysis

6. Conclusion

6.1. Future Directions

7. Acknowledgements

Bibliography

[1] W. H. Guss and R. Salakhutdinov, "On Characterizing the Capacity of Neural Networks using Algebraic Topology," *CoRR*, 2018.

[2] M. Bianchini and F. Scarselli, "On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[3] S. Fitz, "The Shape of Words - topological structure in natural language data ," in *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, in Proceedings of Machine Learning Research, vol. 196. PMLR, 2022, pp. 116–123.

[4] O. Draganov and S. Skiena, "The Shape of Word Embeddings: Quantifying Non-Isometry with Topological Data Analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12080–12099.

[5] S. H. Meirom and O. Bobrowski, "Unsupervised Geometric and Topological Approaches for Cross-Lingual Sentence Representation and Comparison," in *Proceedings of the 7th Workshop on Representation Learning for NLP*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 173–183.

[6] V. Ravishankar, A. Kulmizev, M. Abdou, A. Søgaard, and J. 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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 3031–3045.

[7] L. Kushnareva *et al.*, "Artificial Text Detection via Examining the Topology of Attention Maps," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 635–649.

[8] A. Uchendu and T. Le, "Unveiling Topological Structures from Language: A Comprehensive Survey of Topological Data Analysis Applications in NLP." 2025.

[9] M. R. Costa-jussà *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation ." 2022.

[10] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318.