Assessing the Similarity of Cross-Lingual Seq2Seq Sentence Embeddings Using Low-Resource Spectral Clustering

Nelson Moll
University of Maryland
nmoll@umd.edu

Tahseen Rabbani
Yale University
tahseen.rabbani@yale.edu

Abstract

In this work, we study the cross-lingual distance of machine translations through alignment of seq2seq representations over small corpora. First, we use the M2M100 model to collect sentence-level representations of The Book of Revelation in several languages. We then perform unsupervised manifold alignment (spectral clustering) between these collections of embeddings. As verses between translations are not necessarily aligned, our procedure falls under the challenging, but more realistic non-correspondence regime. The cost function associated with each alignment is used to rank the relative (machine) similarity of one language to another. We then perform correspondent alignment over another cluster of languages, this time using FLORES+ parallel NLLB model embeddings. Our experiments demonstrate that the representations of closely-related languages group closely, and are cheap to align (requiring $<$1000 sentences) via our strategy.

1 Introduction

Assessing the similarities and differences between languages, that is, comparative linguistics, requires the consideration of historical factors, vocabulary, phonology, and written script Georgi et al. (2010); Starostin (2000); Anttila (1989). Computational linguists adopting lexicostatistical techniques can study language distances by measuring the evolution of cognates Gudschinsky (1956). Comparative analysis which operates purely at the word level, such as ranking Levensthein distances (a string-edit metric) Sturrock (2000), has been both widely used and disputed Greenhill (2011). In parallel, the machine learning community recognized the need for sentence-level processing to produce high-quality translations. The attention mechanism, common to transformer-based language models Vaswani (2017), considers the semantic contribution of all tokens (word/sub-word units) in an input to develop an output.

The quality of machine translation has drastically improved in recent years due to the advent of attention-based sequence-to-sequence (seq2seq) models which intake sentences in a source language and output a corresponding translation in a target language Sutskever (2014); Cho (2014). Sharp improvements in multilingual training strategies have resulted in so-called many-to-many translation models that can accept many source-target language pairs. Many-to-many translation models, such as the M2M100 Fan et al. (2021) and NLLB Costa-jussà et al. (2022), can accept pairs from 100 and 200 widely-spoken languages, respectively.

Given a specified source language, the M2M100 and NLLB models tokenize an input and pass it along several attention layers which encode the specified sentence(s) to real-valued embeddings Phuong and Hutter (2022). Such representations produced by deeper encoder layers are thought to embody abstract semantic meaning critical to developing coherent, high-quality output in the decoding phase Vaswani (2017); Clark (2019); Voita et al. (2019). Intuitively, we would expect that closely related languages produce similar representations. If we regard sentences as concepts, language generation benefits from the alignment of closely-related concepts (The et al., 2024). We expect the syntactical and figurative structure of sentences to align more closely among related languages, thus we want to investigate whether many-to-many transformer representations are capturing this dynamic.

In this work, we propose a low-resource strategy for assessing how a many-to-many machine translation model encoder groups languages. First, we collect the sentence representations over a common corpus across a cluster of Slavic, Indo-

Aryan/Dravidian, Romance languages, Scandinavian, Turkic/Mongolic, and Bantu languages. For this paper, we use the mean pooling of hidden states over the entire sequence to get a sentence-level representation, in line with other works Xu et al. (2020); Kudugunta et al. (2019). For one group of language families, the common corpus is the Book of Revelation (BoR). For the other group, the common text is a collection of parallel (i.e., correspondent) sentences from the FLORES-200 dataset Costa-jussà et al. (2022). We validate our method over this correspondent dataset to verify alignment is working as expected in a naive setting. In comparison to the work of Kudugunta et al. (2019) which uses an irreproducible web crawl to generate hundreds of thousands to tens of millions of parallel sentence pairs Uszkoreit et al. (2010), our resource is low-resource: *we only require* < 1000 *sentences per language pair to perform our clustering.*

We treat each language's set of embeddings as a discrete manifold. Then, we perform a pairwise manifold alignment via spectral clustering Wang and Mahadevan (2009) and use the associated cost to produce an ordering of machine-lingual similarities. For the BoR corpus, since translations are not necessarily verse-aligned, we are performing alignment without correspondence – a much more challenging regime, and realistic scenario for ultra low-resource languages. Our similarity rankings over both BoR and FLORES+ closely correspond to established analyses in comparative linguistics Bella et al. (2021) along with a few sharp deviations that may indicate the preference of M2M100 and NLLB to occasionally place representations of less related languages close to one another.

2 Comparison Algorithm

The semantics of a language, referring to its meaning and how words and phrases convey ideas, often follow distinct patterns based on the relationships between words, contexts, and usage. These patterns can be observed in how words group together, how similar meanings emerge in different contexts, or how words with similar meanings are often used in comparable syntactic structures.

Spectral clustering Von Luxburg (2007); Law et al. (2017) can be applied to identify these semantic patterns by analyzing the structure of a similarity matrix constructed from the relationships between words or phrases. We follow the method of Wang and Mahadevan (2009), referred to as manifold alignment without correspondence, and describe this process explicitly in Section 2.1. We must use a "without-correspondence" strategy as variations in translations (in our case, of the Christian Bible), can produce different verse-orderings and shuffled semantics which prevents a verse-to-verse (1-1) correspondence between two languages.

In our case, we used the heat kernel similarity on a suggestion by Wang and Mahadevan (2009) for language comparison. By representing sentences as real-valued vectors in high-dimensional space using encoder embeddings (e.g., M2M100/NLLB representations), we can calculate pairwise similarities, which are then used to create a graph where nodes represent vectors in these representations, and the edges are given quantitatively by their similarity matrix. The spectral clustering algorithm then partitions this graph into clusters by projecting these vector representations onto a set of vectors given by solutions to a generalized eigenvector solution (see Section 2.2). This method hopes to potentially reveal similarities between machine representations of languages by comparing these projections, which are closely related to the clusters. In particular, we examine the square sum of the first d eigenvalues as defined by the general eigenvector equation as given in Section 2.1.

Chowdhury et al. (2021) also used a graph-based approach to study the similarities between languages. They created graph Laplacians between given languages at the word level. Our method considers language at the sentence level and, instead creates a joint graph based on their combined information. Motivated by Wang and Mahadevan (2009), we opt for the combined graph approach due to a belief that we can measure the distance between two languages by considering spectral data associated to a submanifold derived from a combination of data from the graphs of both languages.

2.1 Algorithm Sketch

Let $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \ldots, y_n]$ be $p \times m$ and $q \times n$ matrices, respectively. For our application, X and Y are the mean poolings of hidden sentences states. The rows are the representations and the columns are the features. Let $|| \ x_2 \ ||$ denote the Euclidean distance. Define the $(k+1) \times (k+1)$ matrix R_{x_l} by $R_{x_l}^{i,j} = \frac{\|z_j - z_i\|_2}{\delta_X}$, where $z_1 = x_l$ and $z_2, \ldots, z_{k+1}$ are x_l's k-closest neighbors and δ_X is the standard deviation for the pairwise distances

between the x_j. We now define the similarity matrix W_x by $W_x = \exp\left(-\|R_{x_i} - R_{x_j}\|_F\right)$, where $\|A\|_F = \sqrt{\text{trace}(A^T A)}$ is the Frobenius norm of the matrix A. The matrix W_x is sometimes called the similarity matrix.

Let the diagonal matrix D_x be defined by $D_x^{i,i} = \sum_j W_x^{i,j}$, and let $L_x = D_x - W_x$. We similarly define a family of matrices in terms of Y. Let

$$Z = \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix} \text{ and } D = \begin{bmatrix} D_x & 0 \\ 0 & D_y \end{bmatrix}.$$

Define W by $W^{i,j} = \exp\left(-\text{dist}(R_x, R_y)/\delta_{X,Y}\right)$, where $\text{dist}(R_x, R_y)$ and $\delta_{X,Y}$ are defined in Wang and Mahadevan (2009) or in the Appendix. Let B_x be the diagonal matrix with $B_x^{i,i} = \sum_j W^{i,j}$ and $B_y^{j,j} = \sum_i W^{i,j}$. We define the distance function $d(\cdot)$ as

$$d(R_{x_i}, R_{y_j}) = \min_{1 \le h \le k!} \min\{d_1(h), d_2(h)\}, \text{where}$$
$$d_1(h) = \|\{R_{y_j}\}_h - k_1 R_{x_i}\|_F,$$
$$d_2(h) = \|k_2\{R_{y_j}\}_h - R_{x_i}\|_F,$$
$$k_1 = trace(R_{x_i}^T \{R_{y_j}\}_h)/trace(R_{x_i}^T R_{x_i})$$
$$k_2 = trace(\{R_{y_j}\}_h^T R_{x_i}^T)/trace(\{R_{y_j}\}_h^T \{R_{y_j}\}_h).$$

Here, h is a permutation of the k possible choices for R_{y_i}. The quantity $\delta_{X,Y}$ is the standard deviation of the set $\{dist(R_{x_i}, R_{y_j}) : x_i \in X, y_j \in Y\}$.

Further, define

$$L = \begin{bmatrix} L_x + \mu B_x & -\mu W \\ -\mu W^T & L_y + \mu B_y \end{bmatrix}.$$

Consider the solutions for λ in the equation

$$Z^T L Z \gamma = \lambda Z^T D Z \gamma. \quad (1)$$

Next, index the generalized eigenvalues from least to greatest and consider the first d eigenvalues $\{\lambda_i : 1 \le i \le d\}$ and calculate $K(d) = \sum_{i=1}^{d} |\lambda_i|^2$. This $K(d)$ will be used to measure the alignment quality between two languages.

2.2 Cost Function

The cost function from Wang and Mahadevan (2009) is given as

$$C(\gamma) = C(\alpha, \beta) = \sum_{i,j} \mu(\alpha^T x_i - \beta^T y_j)^2 W^{i,j} + \frac{1}{2} \sum_{i,j} \mu(\alpha^T x_i - \alpha^T y_j)^2 W_x^{i,j} + \frac{1}{2} \sum_{i,j} \mu(\beta^T y_i - \beta^T y_j)^2 = \gamma^T Z^T L Z \gamma,$$

where $\gamma^T = [\alpha^T, \beta^T]^T$ is a solution to the generalized eigenvalue problem Eqn. 1. Note that if we normalize γ by dividing the constant $\sqrt{|\gamma^T Z^T D Z \gamma|}$ then $C(\gamma) = |\lambda|^2$. The cost function $C(\alpha, \beta)$ from Wang and Mahadevan (2009) is minimized by the generalized eigenvectors for the above equation. Hence we define the new cost function $K(d) = \sum_{i=1}^{d} |\lambda_i|^2$, where the λ_i are the eigenvalues for the above equation. The minimum possible value of $K(d)$ is 0 (manifolds are identical) while the maximum is unbounded, though in practice we do not observe it to exceed 1000.

3 Experiments

In this section, we produce a ranking of (machine) language distances. We review our distances against prevailing comparative linguistics theory.

Dataset. Our corpora to compare encoder manifolds are the Book of Revelation (BoR) of the Christian Bible and the FLORES+ dataset (dev split). We choose the BoR due to (1) its widely available translations and (2) since it contains a diverse set of vocabulary and vivid imagery this can help further probe for concept alignment. For the BoR, we source these translations from the digital eBible corpus Akerman et al. (2023). Revelations has a diverse set of words describing abstract visions. We thought this diversity would help separate out some of the differences in the languages we consider. For each family, we attempt to choose translations of the BoR descending from a common pivot or consistent translator, though this is not always possible. FLORES+ sentences are 1-1 aligned between all languages and professionally translated.

Languages. For the non-correspondent BoR clustering task, we consider three clusters of languages: (French, Italian, Spanish, Portuguese), (German, Russian, Ukrainian, Polish), (Kannada, Hindi, Bengali, Gujarati). For the correspondent FLORES+ task, we consider three new clusters of languages: (Icelandic, Swedish, Danish, Norwegian Bokmål), (Swahili, Kirundi, Kinyarwanda, Luganda), (Khalkha Mongolian, Kyrgyz, Tatar, Kazakh). In each quadruplet, we include a challenge (grey) language which is widely accepted to be the most dissimilar of its group despite close geographic proximity.

	Italian	Portuguese	French
Portuguese	239		
French	313	153	
Spanish	101	174	296

Table 1: **Romance Language Distances.** Our method generally places Italian, Spanish, and Portuguese close together, but controversially ranks French closer to Portuguese than the other Romance languages.

	Bengali	Hindi	Kannada
Hindi	21		
Kannada	98	304	
Gujarati	231	360	365

Table 2: **Indo-Aryan/Dravidian Language Distances.** Our ranking overall tends to cluster the Indo-Aryan languages Bengali, Hindi, and Gujarati together. It erroneously places Kannada, a Dravidian language, not as far away for several orderings.

	German	Russian	Polish
Russian	325		
Polish	397	252	
Ukrainian	228	220	155

Table 3: **Slavic/Germanic Language Distances.** Our ranking overall tends to cluster the Slavic languages together.

Model. For each translation of the BoR, we push every verse through the M2M100 (418M model) and extract the mean pooling of hidden states over the entire sequence to get a sentence-level representation. For each language, this results in roughly 403 points in $\mathbb{R}^{1024}$. For translations of FLORES+, we use NLLB (600M model) mean pooling embeddings of 997 sentences also in $\mathbb{R}^{1024}$. We choose $d = 400$ eigenvalues to construct our cost $K(d)$ (as described in Section 2.2 as this explained roughly 90% of covariance across all individual language graph Laplacians. We ran all experiments using only a CPU.

4 Experimental Analysis

4.1 Non-Correspondent Alignment

Tables 1, 2, and 3 depict our seq2seq spectral clustering rankings via manifold alignment without correspondence over the BoR. A higher spectral clustering score indicates a higher cost for manifold alignment.

Our spectral rank successfully tends to group close languages together. This indicates that the manifold alignment is easier for the core similar languages, thus their representations may occupy similar regions in the ambient space. Our ranking, though generally accurate is not immune to errors – for example, placing Kannada, a Dravidian language, very close to some Indo-Aryan languages.

	Swedish	Danish	Nor. Bok.
Danish	299		
Nor. Bok.	76	228	
Icelandic	577	600	619

Table 4: **East/West Scandinavian Language Distances.** Our ranking clusters members of the East Scandinavian family closer together than with Icelandic, which is closer to Old Norse.

	Kh. Mong.	Tatar	Kazakh
Tatar	744		
Kazakh	226	413	
Kyrgyz	600	436	461

Table 5: **Turkic/Mongolic Language Distances.** Kazakh, Tatar, and Kyrgyz (all members of the Turkic Kipchak branch), and generally clustered together. The method commits an error by viewing Khalka Mongolian and Kazakh as closest.

	Swahili	Luganda	Swahili
Luganda	497		
Kirundi	317	298	
Kinyarw.	254	299	282

Table 6: **Great Lakes/Sabaki Bantu Language Distances.** The alignment generally views the Great Lakes Bantu languages as close. Our method commits a single error by viewing Swahili (a Sabaki Bantu language) as the closest language to Kirundi.

4.2 Correspondent Alignment

Tables 4, 5, and 6 depict rankings via manifold alignment with correspondence over FLORES+. To perform parallel alignment, we set $W = I$ in Section 2.1. Our results generally fall in line with what is found in Bella et al. (2021).

Swedish, Danish, and Norwegian Bokmål are closely related members of the East Scandinavian group within the Northern Germanic family and our clustered closely by our method. Kahlkha Mongolian, a member of the Mongolic languages, shares typological features but is less related to the Turkic

group. Our approach does commit an error by judging Khalkha Mongolian as closer to Kazakh than the other Kipchak languages. Swahili, although a Bantu language, is part of the Sabaki group, differs in vocabulary from the other three. Our methodology erroneously views Swahili as the closest language to Kirundi (which is, in fact, Kinyarwanda).

5 Conclusion

In this work, we study how seq2seq translation models group languages together. We conduct this assessment by extracting M2M100 and NLLB hidden representations of sentences of various languages over small, common corpora. We observe that the embedding manifolds of closely related languages likely contain similar structures as they, on average, do not incur high spectral clustering costs. In contrast to Kudugunta et al. (2019), we require < 1000 sentences and can perform clustering without parallel alignment, thus framing our method as a low-resource strategy.

References

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Ramio Anttila. 1989. Historical and comparative linguistics. *John Ben Jamins Publishing Companym*.

Gábor Bella, Khuyagbaatar Batsuren, and Fausto Giunchiglia. 2021. A database and visualization of the similarity of contemporary lexicons. In *International Conference on Text, Speech, and Dialogue*, pages 95–104. Springer.

Kyunghyun Cho. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 375–385.

Kevin Clark. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 385–393.

Simon J Greenhill. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.

Sarah C Gudschinsky. 1956. The abc's of lexicostatistics (glottochronology). *Word*, 12(2):175–210.

Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. 2019. Investigating multilingual nmt representations at scale. *arXiv preprint arXiv:1909.02197*.

Marc T Law, Raquel Urtasun, and Richard S Zemel. 2017. Deep spectral clustering learning. In *International conference on machine learning*, pages 1985–1994. PMLR.

Mary Phuong and Marcus Hutter. 2022. Formal algorithms for transformers. *arXiv preprint arXiv:2207.09238*.

Sergei Starostin. 2000. Comparative-historical linguistics and lexicostatistics. *Time depth in historical linguistics*, 1:223–265.

Shane Sturrock. 2000. Time warps, string edits, and macromolecules–the theory and practice of sequence comparison. david sankoff and joseph kruskal. *Genetics Research*, 76(3):327–329.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

LCM The, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416.

Chang Wang and Sridhar Mahadevan. 2009. Manifold alignment without correspondence. In *IJCAI*, volume 2, page 3.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2020. Probing word translations in the transformer and trading decoder for encoder layers. *arXiv preprint arXiv:2003.09586*.