Exploring Embedding Visualization Tools for Low-Resource Machine Translation

Varin Sikka * and Kevin Dunnell

Posted Date: 13 November 2024

doi: 10.20944/preprints202411.0941.v1

Keywords: LLM Visualization; Low-resource language translation; Latent Lab; Embeddings Visualization

Article

Exploring Embedding Visualization Tools for Low-Resource Machine Translation

Varin Sikka * and Kevin Dunnell

Khan Lab School
MIT
* Correspondence: varins@khanlabschool.org

Abstract: We investigate the potential of embedding visualization tools, specifically Latent Lab, for supporting machine translation tasks with a focus on low-resource languages. Through empirical analysis of bilingual datasets comprising high-resource (Spanish, German, French) and low-resource (Dzongkha) languages, we explore how embedding visualization can facilitate the identification of cross-lingual token correspondences and dataset quality assessment. Our findings suggest that while embedding visualization tools offer promising capabilities for dataset preparation and quality control in machine translation pipelines, several technical modifications are necessary to fully realize their potential. We identify some challenges, and propose concrete modifications to both the datasets and visualization tools to address these limitations.

Keywords: LLM visualization; low-resource language translation; latent lab; embeddings visualization

1. Introduction

The rapid advancement of machine translation has led to impressive coverage across languages, with services like Google Translate now supporting 243 languages. However, significant challenges persist in translation quality [1].

At the same time, the emergence of Large Language Models (LLMs) has raised concerns about their potential negative impact on preserving low-resource languages, as their training data heavily skews toward English [2–4]. Additionally, as LLMs predominantly generate English content, this potentially reinforces existing patterns of English language dominance in digital communication. Understanding these aspects is important for developing more linguistically inclusive AI systems.

Recent initiatives such as Meta's No Language Left Behind (NLLB) [5] and Google's MADLAD [6] have attempted to address this imbalance. However, the fundamental challenge of developing effective machine translation systems for low-resource languages remains.

Embedding visualization tools have the potential to provide powerful ways to explore large unstructured document corpora, and for use in NLP tasks. Tools like Latent Lab [7] enable researchers to explore high-dimensional embedding spaces in an intuitive manner, facilitating the identification of patterns and anomalies in large scale unstructured data. This paper explores how Latent Lab, and similar tools, can support machine translation, and assisting humans in the process, and identifies potential modifications to enhance its capability for this purpose.

2. Our Approach

We propose a novel approach to support machine translation through embedding visualization. We hypothesized that if we embedded English translations (which Latent Lab can process) of sentences in another target language (which Latent Lab cannot process) and used the non-English translation sentences as the titles of the embeddings, we could use Latent Lab to identify feature vectors that link English tokens with their corresponding tokens in the target language. This process could facilitate the training of a translation system by mapping token correspondences between two languages, enabling a novel approach to machine translation.

 doi:10.20944/preprints202411.0941.v1

Figure 1 shows a LatentLab screenshot of the embedding space for English-Dzongkha.

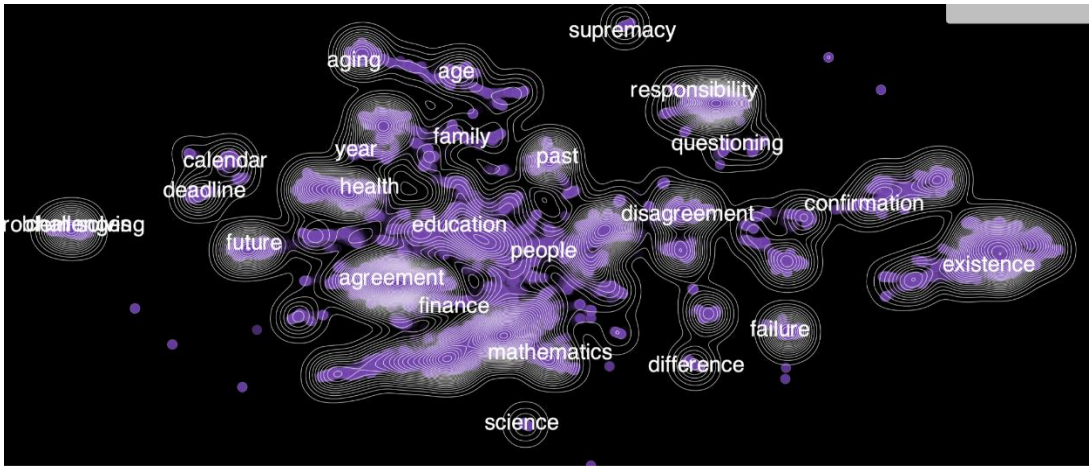

Figure 1. Visual Representation of Dzongkha Dataset on Latent Lab.

2.1. Dataset Selection

To implement our approach, we utilized bilingual datasets from Meta's NLLB corpus, selecting:

- High-resource languages: Spanish, German, French
- Low-resource language: Dzongkha

All four datasets include sentences in the native languages, and their English translations.

2.2. Embedding Generation and Visualization

We used these languages to test whether there is a significant difference in our results between languages closely related and/or similar to English (such as the first three), and those that are not (such as Dzongkha). Then, we uploaded these four datasets to Latent Lab, embedding each English translation.

We processed the datasets using Latent Lab, which leverages the OpenAI API for embedding generation. The tool provides a two-dimensional visualization of the high-dimensional embedding space, enabling identification of clusters and patterns in the data.

3. Results and Analysis

We concluded that Latent Lab and similar tools offer powerful support for cleaning and preparing datasets for machine translation. By allowing users to visualize a large embedding space easily, these tools help identify connections, detect errors, and highlight error-prone areas or deficiencies in the training set. However, our investigation revealed that in its current state, Latent Lab requires several modifications to realize this potential fully. There are several reasons for this:

3.1. Noise in Training Data

We identified significant noise in the datasets, particularly in the Dzongkha corpus. Notable issues included:

- Entries consisting solely of punctuation marks (some of this may have been a kind of word art that was automatically flagged as Dzongkha-language text)
- Presence of text in unrelated languages (e.g., Japanese text in Dzongkha dataset)
- Misclassified non-linguistic content

3.2. Translation Quality

Several categories of translation errors were identified:

- Incomplete translations of longer sentences

 doi:10.20944/preprints202411.0941.v1

- Inconsistent handling of proper nouns. We found that names in one language were replaced with a wholly different and unrelated name in their English translations, for example a German sentence promoting a company having the company name replaced with a different one in the English translation (the sentence "Deshalb haben wir livewire für menschen wie sie geschaffen" was translated as "We built LemonStand for people like you", with the name "livewire" replaced with "LemonStand").
- Grammatically incorrect English translations, such as the sentence "aren't they are they?" in the Dzongkha dataset. Errors like this did not lead to significant errors in Latent Lab's visualization of the embedding space, however, they would cause problems in a future training step as described above, in which case proper noun tokens could get mismatched.

3.3. Distribution Imbalances

Our analysis revealed uneven distribution of content types, particularly in the Dzongkha dataset. The visualization of its dataset in Latent Lab revealed a large cluster of short sentences used to negate a prior sentence in context, such as "And no there isn't," "Oh, there is not," or "No, no it isn't." This was also not a significant issue for Latent Lab at first, but LLMs typically prefer a more even distribution of training data, hence we would expect less clustering like the one mentioned from the Dzongkha dataset.

3.4. Technical Limitations

The current implementation of Latent Lab presents several limitations for machine translation applications:

- Latent Lab presently relies on the OpenAI API to generate embeddings, and we do not have control over the underlying embedding model
- Limited access to full-resolution embedding vectors
- Restricted front-end visualization capabilities

These limitations need to be addressed before proceeding to the next step of associating vector features with linguistic elements.

4. Proposed Improvements

Based on our findings, we propose the following modifications to enhance the effectiveness of embedding visualization tools for machine translation:

1. Dataset Enhancement:
    - Manual construction of larger, cleaner datasets, eliminating noise
    - Comprehensive translation and its validation
    - Balanced content distribution, in particular ensuring that a variety of data is present in the set, and avoiding clusters of similar sentences
2. Technical Modifications:
    - Implementation of full-precision vector access
    - Enhanced visualization capabilities
    - Integration with custom embedding models

5. Conclusion and Future Work

Our investigation demonstrates the potential of embedding visualization tools in supporting machine translation development, particularly for low-resource languages. While current implementations require modifications to fully realize this potential, the approach shows promise for improving dataset quality and facilitating cross-lingual token mapping.

Future Work Should Focus On:

1. Implementing the proposed technical modifications
2. Developing standardized protocols for dataset preparation

 doi:10.20944/preprints202411.0941.v1

3. Evaluating the impact of visualization-guided dataset improvements on translation quality
4. Extending the approach to additional low-resource languages

References

1. Google Translate Community Forum. "Google Translate has changed for the worse but why? Is there any way to fix it?" Available at: support.google.com/translate/thread/276974216
2. World Economic Forum. "Generative AI languages llm" Available at: weforum.org/agenda/2024/05/generative-ai-languages-llm/
3. Guo, Conia, et al. "Do Large Language Models Have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs" arXiv:2410.15956v1
4. Wendler, Veselovsky, et. al. "Do Llamas Work in English? On the Latent Language of Multilingual Transformers" https://arxiv.org/abs/2402.10588
5. OPUS NLLB corpus. Available at: https://opus.nlpl.eu/NLLB/corpus/version/NLLB
6. MADLAD-400 Dataset. Available at: https://huggingface.co/datasets/allenai/MADLAD-400
7. Latent Lab. Available at: latentlab.media.mit.edu