VN-MTEB

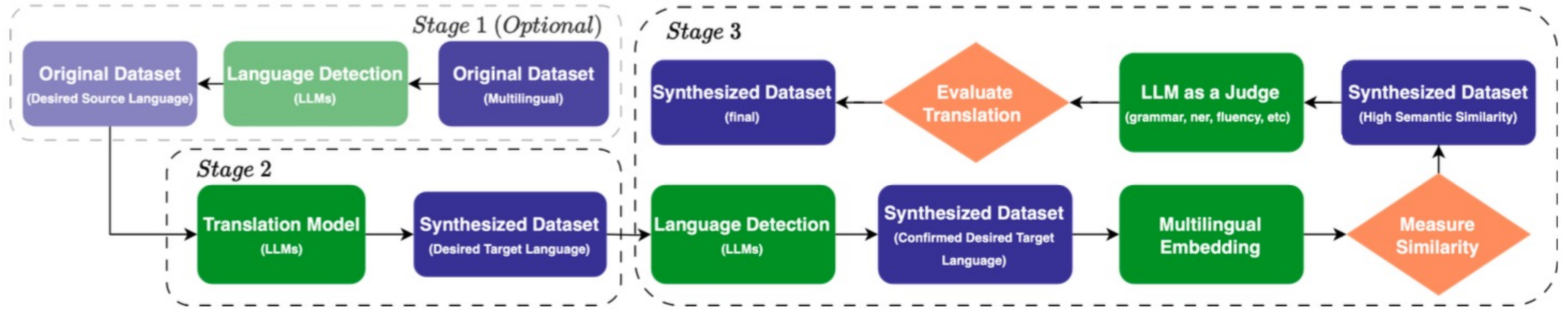
Vietnamese Massive Text Embedding Benchmark

Loc Pham[1] ; Tung Luu[1]; Thu Vo[1]; Minh N.T. Nguyen[2]; Viet Hoang[1]; GreenNodeAI[1]; International University; VNU-HCMC[2]

Abstract

We introduce a Vietnamese benchmark, **VN-MTEB** for embedding models, which we created by translating a large number of English samples from the Massive Text Embedding Benchmark using our new automated framework, thereby contributing an extension of the Massive Multilingual Text Embedding Benchmark with our additional Vietnamese tasks and datasets. We leverage the strengths of large language models (LLMs) and cutting-edge embedding models to conduct translation and filtering processes to retain high-quality samples, guaranteeing a natural flow of language and semantic fidelity while preserving named entity recognition (NER) and code snippets. Our comprehensive benchmark consists of **41 datasets** from six tasks specifically designed for Vietnamese text embeddings. In our analysis, we find that bigger and more complex models using Rotary Positional Embedding outperform those using Absolute Positional Embedding in embedding tasks.

Background and Dataset creation


Stage 1 (Optional)
Original Dataset (Desired Source Language)
Language Detection (LLMs)
Original Dataset (Multilingual)
Stage 2
Translation Model (LLMs)
Synthesized Dataset (Desired Target Language)
Stage 3
Synthesized Dataset (final)
Evaluate Translation
LLM as a Judge (grammar, ner, fluency, etc)
Synthesized Dataset (High Semantic Similarity)
Language Detection (LLMs)
Synthesized Dataset (Confirmed Desired Target Language)
Multilingual Embedding
Measure Similarity

The Figure presents a pipeline for generating a synthesized dataset by transforming a source dataset into a low-resource language through three main stages.

Stage 1, language filtering is performed using an LLM to detect and retain only samples written in the desired source language, ensuring clean input data (this step may be skipped if the entire dataset is translated).

Stage 2, the dataset is translated into Vietnamese using an LLM while preserving semantic meaning, named entities, code snippets, and other important textual features.

Stage 3, the translated data is evaluated through a three-step quality assessment: checking for contamination from other languages, verifying semantic similarity with the original content, and scoring each sample based on multiple evaluation criteria. Samples that do not meet the predefined quality threshold are removed from the final synthesized dataset.

GreenNodeAI - VN-MTEB 6 Tasks - 41 datasets

Retrieval

- ArguAna-VN
- Webis-Touche-VN
- Climate-Fever-VN
- SciFact-VN
- DBPedia-VN
- CQADupstack-VN
- NQ-VN
- HotpotQA-VN
- Trec-Covid-VN
- NFCorpus-VN
- Fever-VN
- Quora-VN
- Scidocs-VN
- Fiqa-VN
- Msmarco-VN
- GreenNodeMarkdownTable

Classification

- AmazonCounterfactual-VN
- AmazonReviews-VN
- AmazonPolarity-VN
- Banking77-VN
- Emotion-VN
- Imdb-VN
- MassiveIntent-VN
- MassiveScenario-VN
- MTOPDomain-VN
- MTOPIntent-VN
- ToxicConversations-VN
- TweetSentimentExtraction-VN

Pair Classification

- SprintDuplicateQuestions-VN
- TwitterSemEval2015-VN
- TwitterURLCorpus-VN

Clustering

- RedditClustering-VN
- RedditClusteringP2P-VN
- StackExchangeClusteringP2P-VN
- StackExchangeClustering-VN
- TwentyNewsgroupsClustering-VN

Reranking

- AskUbuntuDupQuestions-VN
- SciDocsRR-VN
- StackOverflowDupQuestions-VN

Semantic Textual Similarity

- STSBenchmark-VN
- BioSSES-VN
- SICK-R-VN

Results

	Size	Dim	Type	Retr.	Class.	PairClass.	Clust.	Rerank.	STS	Avg. ↑
Num. Datasets (→)	(Params)	(Dim)		15	12	3	5	3	3	41
gte-Qwen2-7B-instruct*	7B	3584	RoPE	**46.05**	70.76	72.09	**53.15**	74.28	78.73	65.84
e5-Mistral-7B-instruct*	7B	4096	RoPE	41.73	72.21	84.01	51.71	**75.15**	81.20	67.67
bge-multilingual-Gemma2*	9B	3584	RoPE	20.52	71.78	66.97	40.13	64.21	66.11	54.95
gte-Qwen2-1.5B-instruct*	1.5B	1536	RoPE	42.01	67.14	72.70	47.64	71.37	79.97	63.47
m-e5-large-instruct*	560M	1024	APE	40.88	**73.39**	**84.47**	52.96	73.28	**82.94**	**67.99**
m-e5-large	560M	1024	APE	37.65	65.03	83.70	45.78	70.40	80.65	63.87
bge-m3	568M	1024	APE	39.84	69.09	84.43	45.90	71.28	78.84	64.90
Vietnamese-Embebedding	568M	1024	APE	34.18	69.06	82.84	45.61	70.89	77.48	63.34
KaLM-embedding-m-mini-v1	494M	896	RoPE	35.07	62.84	79.95	46.85	68.85	78.54	62.02
LaBSE	471M	768	APE	17.77	60.93	77.57	34.59	65.65	72.04	54.76
gte-multilingual-base	305M	768	APE	38.38	64.99	84.42	50.25	71.78	81.51	65.22
m-e5-base	278M	768	APE	34.50	63.29	82.51	45.70	69.07	79.45	62.42
halong-embedding	278M	768	APE	34.45	63.33	81.20	43.42	69.83	77.39	61.60
m-e5-small	118M	384	APE	34.12	60.27	81.18	43.16	67.69	77.56	60.66
vietnamese-bi-encoder	135M	768	APE	25.37	58.92	77.40	34.13	64.95	68.58	54.89
sup-SimCSE-VN-phobert-base	135M	768	APE	12.03	59.69	71.31	33.05	58.86	68.61	50.59
MiniLM-L12	33.4M	384	APE	14.14	45.57	69.46	24.36	60.44	62.34	46.05
MiniLM-L6	22.7M	384	APE	9.65	45.19	66.13	20.40	59.46	58.25	43.18

Average performance of the main metric (in percentage) per task and per model on VN-MTEB subsets. The symbol * indicates that the model is **Instruct-tuned**. Bold values highlight the best results for each specific task. The column "Avg." represents the mean of the average scores across all tasks.

Conclusion

We utilize our proposed translation pipeline for translating 41 datasets from 6 tasks to create a massive text embedding benchmark from English to a low-resource language -Vietnamese. Through ex- tensive experiments on our translation pipeline, we show that with LLMs we can delegate lots of effort from humans to translate a massive dataset with quality. Additionally, we evaluated 18 text embed- dings and revealed the superiority of **RoPE**-based embedding models over **APE**-based ones in some tasks, giving an overview of choices to consider when selecting types of models to put in production and further research.