**Problem**

Semantic Textual Similarity (STS) is a task in which models must score how semantically similar two sentences are. In multilingual STS, the two sentences are in different languages. Achieving strong multilingual STS performance typically requires both parallel corpora (the same sentence translated into multiple languages) and human-annotated sentence pairs (two sentences in different languages with a human annotation). However, this data is often scarce for low-resource languages.

**Example Annotated Sentence Pairs**

- 3/5: A man is playing flute / A man is playing a bamboo flute.
- 0/5: A woman is writing / A woman is swimming

**Background**

An STS model transforms a sentence into a high-dimensional sentence embedding. Following an approach pioneered by Nelles and Gurevych, the predicted semantic similarity of two sentences is the cosine similarity between their sentence embeddings. Our model extends that of Tiyyajam et al. 2021, which introduces a novel architecture to extract language-agnostic semantic embeddings from embeddings produced by pretrained models through contrastive learning.

We apply our model to LaBSE and XLM-R, two large pre-trained multilingual sentence embedding models. These models are pre-aligned, so the same sentence in multiple languages should produce roughly the same embedding.

**Methods**

We extend Tiyyajam et al.’s architecture with a multi-stage training pipeline. This approach allows us to leverage human-annotated sentence data when available for sentence pairs (i.e., for higher-resource languages), while still improving overall STS performance.

- **Stage 1** is a complete reimplementation of Tiyyajam et al. that trains on parallel corpora. We experiment with several extensions to the original architecture.
- **Stage 2** is our own original model, which trains on human-annotated STS data and enhances the embeddings produced by Stage 1. It is a two-layer neural network; we augment in several ways.

![Diagram showing stages 1 and 2 of the model](image)

**Experiments**

To assess general performance—rather than model-specific behavior—we conducted all our experiments on 1008 datasets from both XLM-R and LaBSE.

- **Stage 1:** Takes sentence embeddings from a large multilingual transformer sentence encoder as its input; outputs language embeddings and meaning embeddings. We experiment with adding an additional layer between the transformer embedding model and Tiyyajam et al.'s architecture.
- **Stage 2:** Takes meaning embeddings from Stage 1 and outputs improved meaning embeddings trained on human-annotated data. We experiment with an additional convolutional layer, as well as applying various non-linearities to the similarity scores before comparing to the human STS annotations.

**Analysis**

We found that the best-performing models typically had relatively lower levels of complexity. In particular, the model with the best STS performance on the test set was the simplest Stage 1 model atop XLM-R. Our Stage 2 model (which augments embeddings using human-annotated STS data) improved STS performance on some language pairs, but additional layers of complexity did not, including a fully connected “prelay” in Stage 1, a 1x1 convolutional prelayer in Stage 2, and non-linearities for cosine similarities in Stage 2. These changes did, however, improve performance atop LaBSE across all language pairs.

- Our multi-stage training pipeline improved performance even on held-out language pairs.
- In particular, our best-performing model had higher scores on EN, EN, PT, and PT, STS, indicating that the model was able to learn more general structural patterns in sentence embeddings that were transferable across languages.
- The model’s STS performance improved roughly equally for non-English language pairs as it did for language pairs including English (relative to the baselines).
- The model’s computed similarity scores cluster around 0.75 (seen to the right), while the human scores are uniform, suggesting further STS performance is possible with additional research.

**Conclusions**

We have shown that making use of available human-annotated STS data through a multi-stage training pipeline can lead to improvements in STS performance on state-of-the-art models. The effects of Stage 2 were more pronounced atop LaBSE, where Pearson STS scores on the STS test set were higher across all language pairs. However, Stage 2 still yielded improvements for certain language pairs when using XLM-R as a base, which were notable given XLM-R was already directly trained to achieve high STS performance. Overall, we present a practical approach for STS that enables use of all available data—both parallel corpora and human-annotated STS data—that improves performance even on held-out language pairs.

**References**

- [Bosco, C., Nelles, E., Gurevych, I., (2022)].
- [Nelles, E., Gurevych, I., (2022)].
- [Spear, L. (2010)].

**Acknowledgements**

Thanking Tyiyyajam, Tyiyyajam, and Tyiyyajam, for their invaluable multilingual and cross-lingual feedback and support. In preparation of this submission, we acknowledged the work of our co-authors, in particular Tyiyyajam, for contributions to the methodology and language processing components of this work.