CS224N CUSTOM PROJECT

Lasts4Ever: Language Agnostic Semantic Text Similarity for "Every" Language

Rhythm Garg, Igor Barakaiev, and Miles McCain

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

3.8/5 A man is playing a flute / A man is playing a bamboo flute.
0.5/5 A woman is writing / A woman is swimming
2/5 A person is peeling an onion / A person is peeling an eggplant

From Huggingface's STSb Multi MT. Note that in multilingual STS, the pairs are not in the same language.

Background

An STS model transforms a sentence into a high-dimensional **sentence embedding**. Following an approach pioneered by Reimers and Gurevych, the predicted semantic similarity of two sentences is the **cosine similarity** between their sentence embeddings.

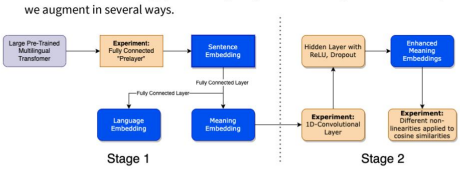
Our model extends that of **Tiyajamorn et al. 2021,** which introduces a novel architecture to extract language-agnostic **meaning embeddings** from embeddings produced by pretrained models through **contrastive learning**.

We apply our model atop **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 Tiyajamorn 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 resourced languages), while still improving overall STS performance.

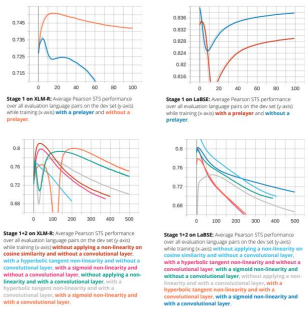
- **Stage 1** is a complete reimplementation of Tiyajamorn et al. that trains on parallel corpora. We experiment with several extensions to the original architecture.
- **Stage 2** is our own original model, which trains on human-annotated STS data and enhances the embeddings produced by Stage 1. It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



Experiments

To assess general performance—rather than model-specific behavior—we conducted all our experiments atop 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 Tiyajamorn 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.

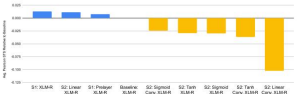


Stage 1 on XLM-R: Average Pearson STS performance over all evaluation language pairs on the dev set (y-axis) while training (x-axis) **with a prelayer** and **without a prelayer.**

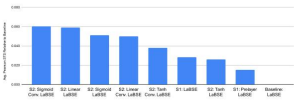
Stage 1 on LaBSE: Average Pearson STS performance over all evaluation language pairs on the dev set (y-axis) while training (x-axis) **with a prelayer** and **without a prelayer.**

Stage 1+2 on XLM-R: Average Pearson STS performance over all evaluation language pairs on the dev set (y-axis) while training (x-axis) **without applying a non-linearity on cosine similarity and without a convolutional layer**, **with a hyperbolic tangent non-linearity and without a convolutional layer**, **with a sigmoid non-linearity and without a convolutional layer**, **without applying a non-linearity and with a convolutional layer**, **with a hyperbolic tangent non-linearity and with a convolutional layer**, **with a sigmoid non-linearity and with a convolutional layer.**

Stage 1+2 on LaBSE: Average Pearson STS performance over all evaluation language pairs on the dev set (y-axis) while training (x-axis) **without applying a non-linearity on cosine similarity and without a convolutional layer**, **with a hyperbolic tangent non-linearity and without a convolutional layer**, **with a sigmoid non-linearity and without a convolutional layer**, without applying a non-linearity and with a convolutional layer, **with a hyperbolic tangent non-linearity and with a convolutional layer**, **with a sigmoid non-linearity and with a convolutional layer.**



XLM-R: Relative average Pearson STS scores on the STSb test set across experiments (n = 1,379). Zero represents XLM-R's baseline performance. S2 (Stage 2) models run atop the S1 (Stage 1) model without a prelayer. Note that EN-EN and EN-PT are held-out (we train our model on no monolingual or PT data).



LaBSE: Relative average Pearson STS scores on the STSb test set across experiments (n = 1,379). Zero represents LaBSE's baseline performance. S2 (Stage 2) models run atop the S1 (Stage 1) model without a prelayer. Note that EN-EN and EN-PT are held-out (we train our model on no monolingual or PT data).

TRAINING DATA

Stage 1

- *Parallel corpora from Tatoeba.com.*
- *Sentences per pair: EN-AR (31,926), EN-DE (395,616), EN-ES (246,850), EN-FR (314,244), EN-IT (558,793), EN-NL (122,305), EN-PL (65,965), EN-RU (619,449), EN-TR (698,829), ES-ZH (10,814), FR-ES (57,097), RU-DE (140,108), TR-PL (935), UA-ES (25,731), ZH-RU (9,657).*
- *Held out for STS evaluation: EN-EN, EN-PT, PT-PL.*
- *Random 80%/20% train-dev split on parallel corpora.*
- *STS performance evaluated using Stage 2 dev set.*

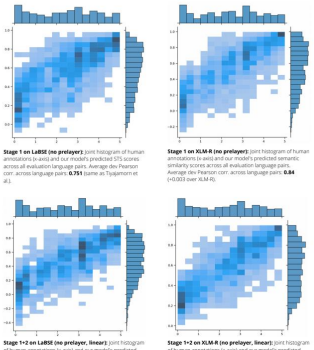
Stage 2

- *Human-annotated STS data (from STSb Multi MT) in EN-DE, EN-ES, EN-FR, EN-IT, EN-NL, NL-IT, and FR-DE.*
- *Held out: EN-EN, EN-PL, EN-PT, EN-RU, RU-DE, FR-ES, ES-ZH, ZH-RU, PT-PL.*
- *5,749 train examples per pair, 1,500 dev examples per pair.*

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 on 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 "prelayer" in Stage 1, a 1-D 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, EN-PT, and PT-PL 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 their 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.



Stage 1 on LaBSE (no prelayer): Joint histogram of human annotations (x-axis) and our model's predicted STS scores across all evaluation language pairs. Average dev Pearson corr. across language pairs: **0.751** (same as Tiyajamorn et al.).

Stage 1 on XLM-R (no prelayer): Joint histogram of human annotations (x-axis) and our model's predicted semantic similarity scores across all evaluation language pairs. Average dev Pearson corr. across language pairs: **0.84** (+0.003 over XLM-R).

Stage 1+2 on LaBSE (no prelayer, linear): Joint histogram of human annotations (x-axis) and our model's predicted semantic similarity scores across all evaluation language pairs. Average Pearson corr. across language pairs: **0.843** (+0.09 over Stage 1).

Stage 1+2 on XLM-R (no prelayer, linear): Joint histogram of human annotations (x-axis) and our model's predicted semantic similarity scores across all evaluation language pairs. Average Pearson corr. across language pairs: **0.843** (+0.003 over Stage 1).

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 STSb 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 was 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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language- agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*. Processing. Association for Computational Linguistics, November 2019.