Contrastive Bilingual Text Embedding Alignment: A Technical Report

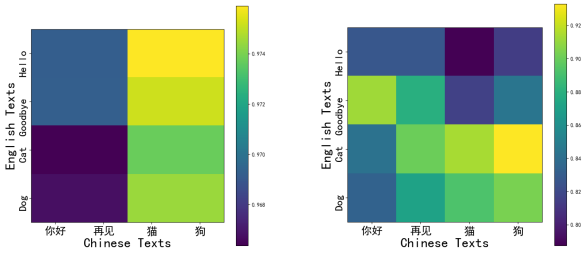
Haoru Li
hal174@ucsd.edu

1 Introduction

Bilingual or multilingual situations are common for international students (like me) and international companies. It is widely known that most modern Natural Language Processing (NLP) applications rely heavily on text embedding to convert natural languages into features that can be handled by machines.

For daily or light applications, usually a pretrained language model will be chosen to produce the text embedding, and one of the most famous and widely used models is BERT (Devlin et al., 2019) and its variants, e.g., RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2020). For more general applications, these models usually have multilingual versions.

So, what do users expect from multilingual text embedding? Intuitively, the sentences from different languages are expected to be projected into the same latent space that different translations share the same embedding. However, does BERT really achieve this?



(a) Vanilla DistilBERT Similarity (b) Multilingual DistilBERT Similarity

Figure 1: Cosine Similarity among most common English words and their Chinese Translations

To take a quick look at the latent space of embeddings produced by DistilBERT, a similarity matrix is computed between the most common English words (Hello, Goodbye, Cat and Dog) and their corresponding Chinese translations according to the [CLS] embedding. The results are shown in Figure 1. It turns out that these embeddings work worse than I thought.

For the base model of DistilBERT, all the similarities shown in Figure 1a are around 0.97 and there is no obvious distinction between the embeddings of words of the same category, for example, the word "Cat" has almost the same similarity with "猫" ("Cat" in Chinese) and "狗" ("Dog" in Chinese), so do the word "Dog". A similar situation also occurs for the word pair "Hello" and "Goodbye". For the multilingual version of DistilBERT, the results shown in Figure 1b still do not meet the expectation. "Hello" has almost the same similarity with "你好" ("Hello" in Chinese), "再见" ("Goodbye" in Chinese) and "Cat" is even more similar to "狗" than "猫".

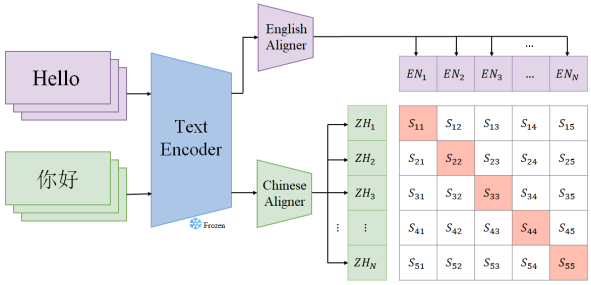


Figure 2: Structure of Contrastive Bilingual Text Embedding Alignment

Therefore, an aligner is needed to align the embeddings between translations. Inspired by CLIP (Radford et al., 2021), a contrastive learning strategy is applied to align embeddings of English and Chinese texts, shown in Figure 2, where the pretrained text encoder is frozen and two MLP aligners are applied to align the English and Chinese embeddings based on contrastive learning.

An English-Chinese translation dataset from WMT2018 (WMT 2018, 2018) is used to align

English and Chinese embeddings. Then, an English-Chinese Biorxiv text clustering dataset is constructed from the original completely English version from Massive Text Embedding Benchmark (MTEB) (MTEB, 2022). The aligned embeddings are tested on the dataset, and a case study is done based on the results.

In summary, this project includes the following tasks:

- Collected and preprocessed English-Chinese translation dataset: DONE.
- Collected and preprocessed English-Chinese text clustering dataset: PARTLY DONE: There is no existing English-Chinese text clustering dataset, so I have to translate on my own. Due to the limited budget on translation API, I just created a relatively small dataset with 1,000 English-Chinese samples.
- Build and train an English-Chinese embedding aligner on English-Chinese translation dataset: DONE.
- Test the performance of aligned embeddings on text clustering task: DONE.

2 Related work

2.1 Text Encoder

BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019) is a powerful and widely used backbone of many NLP applications, so does this project. Here, I chose a distilled variant of BERT, also known as DistilBERT (Sanh et al., 2020), which provides lighter and faster text embedding function compared with original BERT.

2.2 Contrastive Learning

CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) applies contrastive learning to align the embeddings of language and image. CLIP trains two mono-modal encoders for language and image jointly by optimizing the similarity between paired texts and images. By this approach, CLIP pulls the latent spaces of text embedding and image embedding into the same one.

2.3 Multilingual Language Models

There are other approaches to train a multilingual language model. Sentence-BERT (Reimers and Gurevych, 2020) applies a distillation technique that optimizes the similarity between multilingual embeddings produced by the student model and monolingual embeddings produced by the teacher model.

Multilingual E5 (Wang et al., 2024) is another powerful multilingual text encoder by Microsoft. It also utilizes contrastive learning to perform a weakly-supervised pre-training on a massive multilingual dataset, followed by a supervised fine-tuning on high-quality labeled data.

3 Dataset

The dataset consists of two parts: an English-Chinese translation dataset and an English-Chinese text clustering dataset.

The translation dataset is collected by the Conference on Machine Translation (WMT) 2018 (WMT 2018, 2018) that contains 176,943 English and Chinese sentence pairs. I selected 131,072 of them as corpus to align the embeddings of English and Chinese texts. The dataset is divided into an 8:2 ratio for training and validation.

The text clustering dataset is collected from Massive Text Embedding Benchmark (MTEB) (MTEB, 2022) that contains 53,787 Biorxiv titles with their category. To evaluate the embeddings' performance on bilingual text clustering, I pick 1,000 of them and translate the title into Chinese.

3.1 Data preprocessing

Because the backbone text encoder in my project is frozen, to accelerate the training process, I encoded the bilingual translation dataset in advance and saved the embeddings as preprocessed dataset. The training script can directly load corresponding embeddings rather than encode the texts for every epoch.

3.2 Data annotation

To get bilingual English-Chinese text clustering dataset, some samples in the original English text clustering dataset are selected and translated through Google Translation API. However, the free API service is relatively slow, and there are QPS and total character limitations on its use. Due to the limited budget, I only translated 1,000 of the original dataset for a light test.

4 Baselines

One baseline is the original multilingual DistilBERT without alignment and the other baseline is

multilingual DistilUSE, a distilled version of multilingual Universal Sentence Encoder (USE).

The dimension of embeddings produced by multilingual DistilBERT is 768 and the dimension of embeddings produced by multilingual DistilUSE is 512.

The baseline clustering algorithm used for evaluation is K-Means whose K is set to the number of clusters in the test dataset. No other additional hyperparameter needed. I will also dive into the similarity among embeddings of some selected bilingual samples.

5 My approach

5.1 Contrastive Bilingual Text Embedding Alignment

As shown in Figure 2, the text encoder in my approach is frozen, and the trainable part is two aligner of English and Chinese texts.

The structure of aligner is relatively very simple as they are just two n-layer MLPs, where n is a hyperparameter to adjust the complexity of the aligner. I did not apply attention-based structure because the input of aligner is only the embedding of [CLS] token rather than the embeddings of all tokens in the sentence. (This can be a further research for a more powerful alignment)

I applied NT-Xent (Normalized Temperature-scaled Cross-Entropy) Loss, a specified version of InfoNCE (Information Noise Contrastive Estimation) Loss in contrastive learning. NT-Xent Loss computes the cosine similarity between each in-batch embeddings and involves a temperature parameter τ to scale the difference in softmax function. The tow aligners will try to maximize the cosine similarity between English and Chinese embeddings from the same pair and prohibits the similarity with embeddings from other pairs.

5.2 Implementation

The key component of aligner is implemented in `model.py`, including the class for bilingual embeddings dataset `BilingualEmbeddingDataset` that will preprocess the text into embeddings and save to the desk, repetitive n-layer MLP class `MLP`, aligner class `CLIPAlign` and loss function `nt_xent_loss`.

The training script is `train.py`. It will produce a log file when executed, including the training and validation loss, training and validation contrastive accuracy and test adjust rand index of every epoch. Notice that even though I run the test after every epoch, no hyperparameter or model checkpoint is selected according to the test results. The test is executed to observe the relationship between test performance and other validation metrics. The final model is selected with the maximum validation contrastive accuracy.

To test different baselines on English / English-Chinese / Chinese text clustering dataset, run `baseline.py`. System arguments are required to specify the baseline model and dataset.

5.3 Hardware Configuration

The training script is running on my personal laptop, and the hardware configuration is as follows:

- CPU: AMD Ryzen 7 6800H with Radeon Graphics 3.20 GHz
- GPU: NVIDIA GeForce RTX 3070 Ti Laptop

5.4 Runtime

The run time varies according to different complexity of aligners, shown in Table 1. Assuming the aligner's input dimension is D_i and output dimension is D_o, and the hidden dimension of repetitive internal hidden layer is D_h. Hyperparameter aligner complexity N_l defines the number of hidden layers in the aligner. To be more specific, when $N_l = -1$, the aligner is only a linear layer with given D_i, D_o; when $N_l = 0$, the aligner contains two linear layers of $D_i \times D_h$ (input layer) and $D_h \times D_o$ (output layer); when $N_l \geq 1$, the aligner will include N_l $D_h \times D_h$ linear layers between input and output layer.

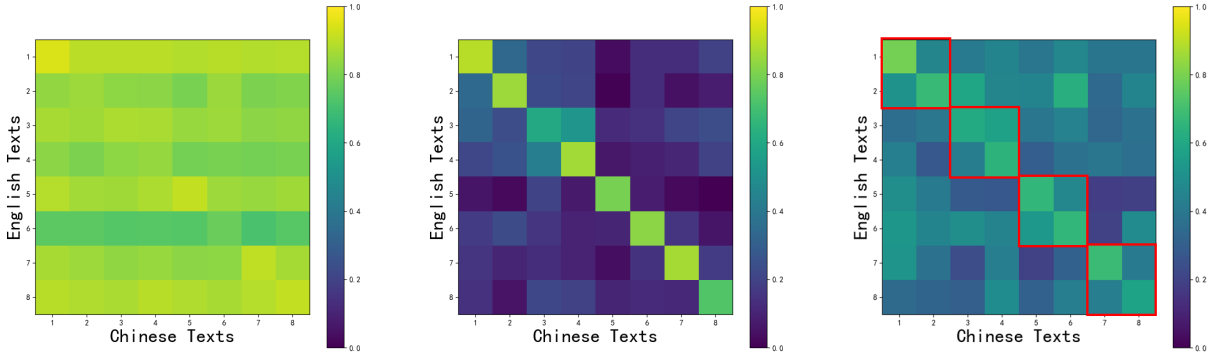
Table 1: Run Time of 50 Epochs with different complexity

Complexity	-1	0	1	2
Run Time (s)	297.21	417.22	532.70	681.12

5.5 Results

5.5.1 Similarity

During validation, the aligner shows low loss and high contrastive accuracy, which means the aligner optimizes the similarity between corresponding English and Chinese texts well. To visualize the results, I picked English and Chinese versions of 2 samples from 4 categories from the test dataset,



(a) Raw Multilingual DistilBERT Embeddings

(b) Raw Multilingual DistilUSE Embeddings

(c) Aligned Multilingual DistilBERT Embeddings

Figure 3: Cosine Similarity among English Biorxiv Titles and their Chinese Translations

and inspected the similarity among 8 English titles and their corresponding Chinese translations, shown in Figure 3.

It can be observed that the similarity among raw embeddings are almost uniform for each English title, and there is no obvious highlight along the diagonal. However, the similarity among aligned embeddings, and the similarity between samples from the same category (outlined in red) are higher than similarities with samples from other categories. In comparison, embeddings produced by DistilUSE exhibits strong similarity between corresponding translations, but weak similarity between samples from the same category.

5.5.2 Adjusted Random Index

Adjusted Random Index (ARI) is a commonly used clustering quality evaluation indicator that can measure the consistency between clustering results and true labels and is insensitive to label exchange. ARI is applied to evaluate the results of clustering using embeddings produced by different models. The test ARI of two baselines and aligner with different complexity is shown in Table 2.

The results indicate that even though aligner greatly reduced the gap between English and Chinese embeddings, the ARI of clustering decreased rather than increased, which is quite out of expectation.

6 Error analysis

The failure of achieving higher ARI by aligning bilingual embeddings may be caused by the loss of semantic information, because as the complexity of aligner increases, the test ARI drops quickly. The linear transform of aligner may disrupt the original information and add some noise, leading to worse clustering performance.

In contrast, the distillation method used by DistilUSE involves consistent guidance from teacher model to make sure the student model keeping semantic information across different languages. While the contrastive learning of my approach does not involve any guidance related to semantic understanding, the model just focused on maximizing the similarity of corresponding bilingual embeddings. That's why multilingual E5 applies further fine-tuning on high-quality labeled data after contrastive pre-training.

Another potential problem is that the backbone language models do not have enough knowledge to catch the biological concept in these titles. Even the best-performed baseline DistilUSE only reaches an ARI less than 0.1, which means a relatively bad performance. I also tried to search for another text clustering dataset, but it is hard to find

Table 2: Adjusted Random Index of Clustering

Model	DistilUSE	DistilBERT	Aligner -1	Aligner 0	Aligner 1	Aligner 2
Test ARI	0.0774	0.0317	0.0224	0.0192	0.0166	0.0079
Best Valid ARI Acc	-	-	0.0256	0.0237	0.0240	0.0146
Best Valid ContraAcc	-	-	0.9823	0.9623	0.9432	0.8537

an ideal one.

I did a lot analysis by organizing data manually like Figure 3c and tried to find some pattern of the clustering results of baseline models and different aligners, but no obvious pattern can be found. I have to say that I lack biological knowledge just like these models. At least from the angel of sole text similarity, aligned embeddings definitely outperforms raw ones. The predicted labels of different aligner is included in the submitted folder `results`.

7 Conclusion

The contrastive learning approach successfully aligns the embedding from two different languages with greatly improved corresponding similarity and contrastive accuracy. This is a relatively efficient and light method to align multilingual embeddings. However, the aligned embeddings did not gain any improvements on text clustering task. This may be caused by the lack of professional domain knowledge and semantic guidance during training.

If there are further chances to work on this project. I want to do a more comprehensive evaluation on the performance of aligned embeddings, involving more downstream NLP tasks like multilingual classification, machine translation and so on.

Moreover, I want to improve the training framework to introduce supervised training enhancing the semantic information in aligned embeddings.

8 Acknowledgements

Great thanks to Prof. Ndapa Nakashole! CSE 256 is the most valuable course I've ever take (I was even asked about some questions in the lectures in my interview). The coding projects and this final project are valuable hands-on experience of NLP.

Also, thanks to the teaching assistants answering my questions and grading my projects!

This report is **completely written by myself** without the assistance from Gen AI.

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

MTEB (2022). mteb/biorxiv-clustering-s2s. https://huggingface.co/datasets/mteb/biorxiv-clustering-s2s. [Accessed: Nov. 1, 2018].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report.

WMT 2018 (2018). THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18). https://www.statmt.org/wmt18/translation-task.html. [Accessed: Nov. 1, 2018].