Enhancing Context Through Contrast

Kshitij Ambilduke Aneesh Shetye Diksha Bagade Rishika Bhagwatkar Khurshed Fitter
Prasad Vagdargi Shital Chiddarwar

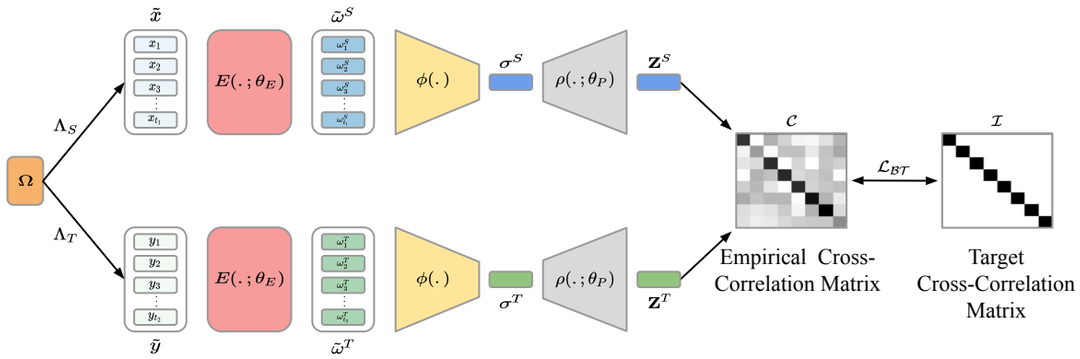
Abstract

Neural Machine Translation (NMT) benefits from semantically rich representations. Considerable progress in learning such representations has been achieved by Language Modelling (LM) and Mutual Information (MI) maximization objectives using Contrastive Learning (CL). The language-dependent nature of LM introduces a trade-off between the universality of the learned representations and the model's performance on the LM tasks. Although CL improves performance, its success cannot be attributed to MI alone. We propose a novel Context Enhancement step to improve performance on NMT by maximizing MI using the Barlow Twins loss. Further, we do not explicitly augment the data but view languages as implicit augmentations, eradicating the risk of disrupting semantic information. Finally, our method does not learn embeddings from scratch and can be generalised to any set of pre-trained embeddings.

Motivation

- We aim to remove language dependent information from the sentence embedding of the source sentence to improve performance on NMT.
- Augmentations in sentences distort their semantics. Hence unlike previous approaches, we view parallel sentences as augmentations of abstract meaning.
- Nor do we depend upon any other pre-training objectives like MMLM (Devlin et al, 2019) and TLM (Lample et. al, 2019) as they require the model to learn language specific information.
- Recent works show that the success of contrastive losses maximizing the lower bound of mutual information does not depend on MI alone.

Approach

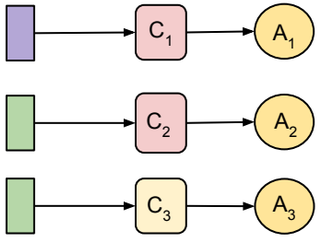


- We intend to improve NMT performance by maximizing the mutual information and minimizing the redundant language-specific information between representations of pairs of sentences of a parallel corpus.
- To achieve this we use an instantiation of the Information Bottleneck Principle (Tishby et al, 2015) through Barlow Twins (Jure, et al, 2021), and do not directly maximise any lower bound estimates on the MI unlike previous approaches.
- We do not learn embeddings from scratch, hence our method and experiments can be generalised to any set of pre-trained embeddings.

Loss function

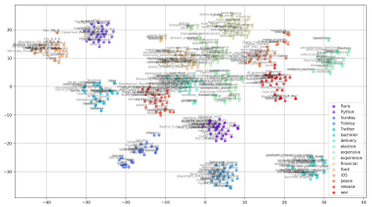
$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}} \qquad \mathcal{C}_{ij} \triangleq \frac{\sum_b z_{b,i}^S z_{b,j}^T}{\sqrt{\sum_b (z_{b,i}^S)^2} \sqrt{\sum_b (z_{b,j}^T)^2}}$$

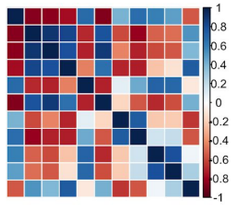
Experiments



To evaluate the language-agnosticism of the embeddings learned by our model, we perform language classification on them.

We visualize the t-SNE plot of word and sentence embeddings to understand their distribution.





To understand the effect of redundancy reduction at sentence level, we plot the correlation matrices between corresponding word pairs at different stages of the CE step.

Datasets

- **WMT 2014 English-German**
- **WMT 2014 English-French**

(We will expand our evaluation to other language pairs from distant families following preliminary results.)

References

Image Source: https://bit.ly/3rh50EE (t-SNE Plot); https://bit.ly/3p5ZKBc (Cross-correlation matrix)