A Python Tool for Selecting Domain-Specific Data in Machine Translation

Pourmostafa Roshan Sharami, Javad; Shterionov, Dimitar; Spronck, Pieter

Published in:
Proceedings of the 1st Workshop on Open Community-Driven Machine Translation

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Pourmostafa Roshan Sharami, J., Shterionov, D., & Spronck, P. (2023). A Python Tool for Selecting Domain-Specific Data in Machine Translation. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation* (pp. 35-36). Universitat d'Alacant. https://hplt-project.org/static/media/proceedings.37b7e88ce3dbab99adf9.pdf#page=35

A Python Tool for Selecting Domain-Specific Data in Machine Translation

Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck
Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands
`{j.pourmostafa,d.shterionov,p.spronck}@tilburguniversity.edu`

1 Introduction

As the volume of data for Machine Translation (MT) grows, the need for models that can perform well in specific use cases, like patent and medical translations, becomes increasingly important. Unfortunately, generic models do not work well in such cases, as they often fail to handle domain-specific style and terminology. Only using datasets that cover domains similar to the target domain to train MT systems can effectively lead to high translation quality (for a domain-specific use-case) (Wang et al., 2017; Pourmostafa Roshan Sharami et al., 2021; Pourmostafa Roshan Sharami et al., 2022). This highlights the limitation of data-driven MT when trained on general-domain data, regardless of dataset size.

To address this challenge, researchers have implemented various strategies to improve domain-specific translation using Domain Adaptation (DA) methods (Saunders, 2022; Sharami et al., 2023). The DA process involves initially training a generic model, which is then fine-tuned using a domain-specific dataset (Chu and Wang, 2018). One approach to generating a domain-specific dataset is to select similar data from generic corpora for a specific language pair, and then utilize both general (to train) and domain-specific (to fine-tune) parallel corpora for MT. In line with this approach, we developed *a language-agnostic Python tool implementing the methodology proposed by Sharami et al. (2022)*. This tool uses monolingual domain-specific corpora to generate a parallel in-domain corpus, facilitating data selection for DA.

The tool's operation requires three inputs: (i) a parallel generic corpus for the source language, (ii) a parallel generic corpus for the target language (iii) a monolingual domain-specific corpus for the source language. Additionally, users can input their desired number of selected data as an optional parameter. Once these inputs are provided, the pre-trained S-BERT (Reimers and Gurevych, 2019) model is employed to transform inputs (i) and (iii) using Siamese and triplet networks. We reduced the original word embedding dimension from 768 to 32 using PCA (Jolliffe, 2011) to make it less computationally expensive. If the size of the corpus (iii) is exceeded by the desired number of selected data, the generic corpora are split into multiple equal parts, and each of these parts is used separately in the subsequent step.

The final step involves using semantic search to find generic sentences that are similar to domain-specific data. This is done by comparing the vectors of sentences and ranking them based on their cosine similarity score. The sentence with the highest similarity score is labeled as Top 1, while the one with the lowest similarity score is labeled as Top N. The default value for N is 5, which is based on the original research paper, but users can choose a different value for N. For each split, the tool then creates a CSV file that includes information about the domain-specific sentence (labeled as `Query`), the top selected source and target sentences (labeled as $topN_{src}$ and $topN_{trg}$), and their corresponding similarity scores. By concatenating the CSV columns generated, one can obtain as much data as previously requested.

Our tool is particularly useful to the MT community as it addresses the scarcity of parallel domain-specific data across different language pairs. By using our tool, users can seamlessly select domain-specific data from generic corpora to train a domain-specific MT model. This tool is typically used when there is a lack of domain-specific data or when only monolingual data is available. However, our tool is generic and not limited to the size of the domain-specific data.

Our tool is licensed under the MIT License and is accessible to the public for free at `https://github.com/JoyeBright/DataSelection-NMT/tree/main/Tools_DS`.

References

Chu, Chenhui and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jolliffe, Ian, 2011. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg.

Pourmostafa Roshan Sharami, Javad, Dimitar Shterionov, and Pieter Spronck. 2021. A novel pipeline for domain detection and selecting in-domain sentences in machine translation systems. In *The 31st Meeting of Computational Linguistics in The Netherlands (CLIN 31)*.

Pourmostafa Roshan Sharami, Javad, Elena Murgolo, and Dimitar Shterionov. 2022. Quality estimation for the translation industry – data challenges. June. The 32nd Meeting of Computational Linguistics in The Netherlands, CLIN ; Conference date: 17-06-2022 Through 17-06-2022.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Saunders, Danielle. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Sharami, Javad Pourmostafa Roshan, Dimitar Shterionov, and Pieter Spronck. 2022. Selecting parallel in-domain sentences for neural machine translation using monolingual texts.

Sharami, Javad Pourmostafa Roshan, Dimitar Shterionov, Frédéric Blain, Eva Vanmassenhove, Mirella De Sisto, Chris Emmery, and Pieter Spronck. 2023. Tailoring domain adaptation for machine translation quality estimation.

Wang, Rui, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada, July. Association for Computational Linguistics.