Data Efficient Machine Learning for Educational Content Creation

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

Ayush Maheshwari
IIT Bombay
ayusham@cse.iitb.ac.in

ABSTRACT

Machine Learning has revolutionized education by offering numerous practical applications. One such application is Neural Machine Translation (NMT) systems in the field of education, which hold immense social importance. These systems have the potential to make information accessible to diverse users in multilingual societies. By effectively translating audio, video, and textual content into vernacular languages, NMT systems greatly assist both students and teachers. However, when it comes to translating higher education or technical textbooks and courses, it becomes crucial for MT systems to adhere closely to the specific lexicon of the source and target domains. In this tutorial, we present our approach and framework to enable domain-aware translation without the need for parallel domain corpus. We will demonstrate several use-cases and applications that is widely used by hundreds of translators. We will also present our post-editing tool that assists translators in quickly correcting the machine translated text and reduce the cognitive load of users.

Keywords

neural machine translation, OCR, dictionary generation, human-in-the-loop learning, data efficient machine learning

1. INTRODUCTION

The field of Neural Machine Translation (NMT) has achieved remarkable success in achieving state-of-the-art translation capabilities across various language pairs [1]. However, in domain-specific scenarios such as technical content translation, the generic NMT pipeline falls short in guaranteeing the inclusion of specific terms in the translation output. Inclusion of a pre-specified vocabulary becomes crucial for ensuring practical and reliable machine translation (MT). While incorporating domain-specific terms has been relatively easier in phrase-based statistical MT, it poses a challenge in NMT due to the complexity of directly manipulating output representations from the decoder [8]. As an alternative, domain-specific NMT systems have been proposed to generate translations that are aware of the domain by fine-tuning generic NMT models using domain-specific parallel text. However, this approach requires curating translation pairs for each domain, which demands significant human effort and increases the cost of maintaining separate models for each domain. Therefore, it is essential for the MT output to adhere to the source domain by adopting domain-specific terminology, thus reducing and potentially guiding the post-editing effort in translation.

To address this issue, lexically constrained techniques have been employed in NMT, incorporating pre-specified words and phrases in the translation output [4, 3, 2]. In addition to the source sentence, word or phrasal constraints in the target language are provided as input. These constraints can be derived from in-domain source-target dictionaries or can be user-provided source-target constraints during interactive machine translation. Often, these constraints may encode multiple potential translations for a given source phrase. For example, the word 'speed' can be translated into 5 different Hindi phrases *teja, dauḍa, gati, raphtār, cāla* in the physics domain. However, existing constrained translation approaches do not accommodate such ambiguity in the constraints.

2. IMPACT OF THE WORK

The project `https://udaanproject.org` is an end-to-end Machine Translation Framework that includes extensive use of OCR, lexical resources, data efficient learning (open sourced at `https://decile.org`) and a human-in-the-loop machine learning based post-editing platform. This project is an outcome of our Data Efficient Machine Learning[5, 6] (`https://decile.org`) and Natural Language Processing from our group at IIT Bombay. The Udaan project is being used extensively by several, including AICTE (`https://www.aicte-india.org`) for speedy translation of 100s of textbooks into multiple Indian languages. MoUs are also being signed with several state governments - Govt of Maharashtra entered into agreement for usage of `https://udaanproject.org` in the presence of Governor, Education Minister and Director IITB (see `https://udaanproject.org/MediaCoverage?type=mou`)

In this tutorial, we provide insights from our translation ecosystem (https://udaanproject.org) that has helped in translating 100s of diploma and engineering books each in more than 11 Indian languages. We will provide the audience with

G. Ramakrishnan and A. Maheshwari. Data efficient machine learning for educational content creation. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 585–587, Bengaluru, India, July 2023. International Educational Data Mining Society.

© 2023 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
`https://doi.org/10.5281/zenodo.8115780`

the holistic view of:

1. How to build a domain-specific lexicon in 11 Indian languages using a small seed dictionary by utilising the innate connection across languages
2. How to build an multilingual NMT model that ingests domain-specific lexicon without affecting the fluency of the predicted sentence
3. How to build a human-in-the-loop AI post-editing tool that benefits from complex OCR (Optical character recognition) and layout analysis to preserve bounding boxes in the source document. and that learns from the user edits and calibrates the output for subsequent occurrences.
4. What are the insights gathered from translating sample 50 books across 11 languages?

The ecosystem at `https://udaanproject.org` that will be presented as a tutorial is fueled by several peer reviewed publications (`https://udaanproject.org/Publications`).

3. UDAAN POST EDITING TOOL

We present UDAAN, an open-source post-editing tool designed to streamline the manual editing process and facilitate the production of high-quality documents in multiple Indic languages[7]. Post-editing lengthy documents that have been translated is often a laborious task, as editors face difficulties in maintaining consistency between the translated and source texts within the document. Existing tools, although retaining source-target user edits through translation memory (TM), fail to provide consistent suggestions throughout the document.

UDAAN offers an end-to-end Machine Translation (MT) plus post-editing pipeline, allowing users to upload a document and obtain raw MT output. Subsequently, users can utilize our tool to edit the raw translations. UDAAN incorporates several advantageous features:

1. Domain-aware, vocabulary-based lexical constrained MT.
2. Source-target and target-target lexicon suggestions for users, employing lexicon alignment between the source and target texts for replacements.
3. Translation suggestions based on user interaction logs.
4. Source-target sentence alignment visualization, reducing cognitive load during the editing process.
5. Translated outputs available in multiple formats, including docs, LaTeX, and PDF.

Our tool offers several advantages: Firstly, it generates domain-aware raw MT by applying lexical constraints to the translations using domain-specific vocabulary. Secondly, users can incorporate lexicons from both the source-target language and the target-target language. Lexicon-based replacements are determined through alignment between the source and target texts. Additionally, the tool continuously records target-target edits made by users, which can be utilized as suggestions within the tool. Thirdly, the tool leverages user edits to improve translation suggestions. Fourthly, the rich text editor of UDAAN includes sentence alignment visualization between the source and target texts, simplifying the editing process and reducing cognitive load. Lastly, users can download the output document in various formats, including docx, LaTeX, and PDF.

Furthermore, UDAAN provides access to approximately 100 in-domain dictionaries to facilitate lexicon-aware machine translation. Although our experiments are limited to English-to-Hindi translation, the tool is language-agnostic. Based on user feedback and experimental results, UDAAN has demonstrated a significant reduction in translation time, approximately three times faster than the baseline method of translating documents from scratch. UDAAN is available for both Windows and Linux platforms, with its source code accessible on our website at Our tool is available for both Windows and Linux platforms. The tool is open-source under MIT license, and the source code can be accessed from our website, `https://www.udaanproject.org`. Demonstration and tutorial videos for various features of our tool can be accessed here. Our MT pipeline can be accessed at `https://udaaniitb.aicte-india.org/udaan/translate/`.

3.1 Acknowledgments

Ayush Maheshwari is supported by a Fellowship from Ekal Foundation (`www.ekal.org`). Ganesh Ramakrishnan is grateful to NLTM OCR Bhashini project as well as the IIT Bombay Institute Chair Professorship for their support and sponsorship.

4. REFERENCES

[1] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, and M. a. H. Huck. Findings of the 2019 conference on machine translation (WMT19). page 9. Frontiers, 2018.

[2] G. Chen, Y. Chen, Y. Wang, and V. O. Li. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3587–3593, 2021.

[3] G. Dinu, P. Mathur, M. Federico, and Y. Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, 2019.

[4] C. Hokamp and Q. Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, 2017.

[5] A. Maheshwari, O. Chatterjee, K. Killamsetty, G. Ramakrishnan, and R. Iyer. Semi-supervised data programming with subset selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4640–4651, 2021.

[6] A. Maheshwari, K. Killamsetty, G. Ramakrishnan, R. Iyer, M. Danilevsky, and L. Popa. Learning to

robustly aggregate labeling functions for semi-supervised data programming. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1188–1202, 2022.

[7] A. Maheshwari, A. Ravindran, V. Subramanian, and G. Ramakrishnan. Udaan-machine learning based post-editing tool for document translation. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 263–267, 2023.

[8] R. H. Susanto, S. Chollampatt, and L. Tan. Lexically constrained neural machine translation with levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, 2020.