pangeanic

Pangeanic Oct2024

Guidelines and Best Practices for MT Data Curation

V1.0

Pangeanic Oct2024

AI4C PECAT emerged from the need for effective data cleaning in the context of the EuropeanaTranslate project, which aimed to enrich and translate Europeana's multilingual cultural heritage collections. Data cleaning is particularly important for Cultural Heritage professionals who serve as data providers, ensuring that the metadata they contribute (e.g., titles, descriptions, creator information, etc.) is accurate and ready for translation.

The EuropeanaTranslate project focused on translating automatically metadata to make European cultural heritage more accessible in multiple languages.

This document provides a collection of best practices and recommendations for data curation in the training of machine translation (MT) models.

1. How an MT (Machine Translation) system works

A Machine Translation (MT) system uses various algorithms and techniques to analyze the input text in the source language and generate the corresponding translated text in the target language. Frequently, this system core is a neural network that learns from the data it has been trained with.

A general overview of how an MT system works can be found below:

1. Pre-processing: The input text is preprocessed, which involves tasks such as sentence segmentation, tokenization (breaking the text into words or smaller units), and natural language processes which help the neural network to understand better the input text (processes such as tags handling in documents, punctuation, capitalization normalization, etc.).
2. Translation: The system generates the translation by selecting appropriate words, phrases, or sentences in the target language.
3. Post-processing: The generated translation undergoes post-processing, which involves tasks like reordering words or phrases to match the target language conventions and making the translation more fluent and natural.
4. Output: The final translated text is presented to the user as a written translation.

2. Identifying Data Quality: Distinguishing between reliable and unreliable data

The EuropeanaTranslate project involved the fine-tuning of machine translation (MT) models developed under the NTEU project for the European Commission (EC). During the NTEU project, 506 neural machine translation engines were created, covering translations between all EC languages.

In the AI4Culture project, the EuropeanaTranslate engines, specifically those translating from any language into English, were fine-tuned to adapt them to the cultural heritage domain. For this task, various Cultural Heritage partners provided both monolingual and bilingual textual data. However, the provided data often had quality issues that needed to be addressed, both **manually** and through **automated** processes, to ensure the effectiveness of the fine-tuning.

3. Fine-tuning

Using reliable data during the fine-tuning process of an AI model is of utmost importance to guarantee precise predictions. As the model learns from the data it is provided, utilizing low-quality data will result in correspondingly subpar predictions.

"Rubbish in, rubbish out"

4. In-domain data

To achieve improved results within a specific domain, it is important to train models using specialized data. Attempting to obtain specialized results by incorporating generic data into a generic model is ineffective. Hence, the chosen data should consist of domain-specific segments, while minimizing the inclusion of generic domain segments.

5. Generalization

An AI model trained on reliable data is more likely to generalize well to real-world scenarios. Reliable data reflects the true distribution, and patterns present in the target domain, including the different possible polysemy instances.

6. Alignment problems

In certain cases, files containing bilingual data are not correctly aligned, which can lead to numerous issues if left unresolved before the training of the model starts.

7. Good practices for Data Curation

1. **Avoid using repetitions for training**

Training a machine translation (MT) model with repeated sentences can be considered ineffective or useless for distinct reasons:

Redundancy: Repeated sentences provide redundant information to the model. If the same sentence occurs multiple times in the training data, the model may simply memorize the repetitions instead of learning the underlying patterns and generalizing from them.
Bias: Repeated sentences can lead to biased representation of the data. If a particular sentence is excessively repeated, it may dominate the training process and skew the model's understanding of the overall distribution of sentences. Consequently, the model may exhibit a biased translation performance and struggle with diverse or uncommon sentence patterns.
Limited Vocabulary: Training a model on repeated sentences may result in the model learning translations for those specific sentences. As a result, when encountering new or uncommon phrases, the model may struggle to generate accurate translations or provide suitable alternatives.
Overfitting: Overfitting occurs when a model becomes too specialized in the training data and fails to generalize well to unseen data. Repeated sentences may cause the model to over-emphasize specific examples, making it less capable of translating diverse and unseen sentences accurately.

To ensure effective training of an MT model, it is crucial to curate a diverse and balanced dataset that covers various sentence structures, vocabulary, and domains. By exposing the model to a wide range of linguistic patterns, it can better generalize and provide accurate translations for new and unseen sentences.

2. **Use relevant data**

The selected base model is usually trained on a vast amount of generic data. However, when the fine-tuning technique is used, it is crucial to carefully select data that includes vocabulary and information relevant to their specific goals (e.g. adapting the model to a specific domain, industry, company, etc.). Without this targeted selection, the resulting translations may not show significant improvements.

8. Fine-tuning Limitations

While performing a fine-tuning of an MT model can improve the output results by offering an adapted translation, there are still some limitations:

- Data Bias: To get a customized performance, data scientists need to train the model with their own data, which is usually created in-domain. Using an extremely specific type of data (only representative of the target domain or language pairs) can lead to biased translations or poor performance on certain inputs.
- Out-of-Domain Translation gets bad results: If the input falls outside the domain the model was fine-tuned on, its performance may degrade. Fine-tuning may not effectively handle translation requests for highly specialized domains.
- Limited Context Understanding: MT models, including fine-tuned ones, often struggle with capturing and understanding the context of the input. They may produce translations that are technically correct but fail to convey the intended meaning or nuance, especially in cases where the input relies on broader context or cultural references.
- Rare or Unseen Words: Fine-tuned models might not handle rare or unseen words well, especially if the training data did not contain enough instances of such words. These models heavily rely on the vocabulary they were trained in and may struggle with accurately translating words that are not present in their training data.

9. How to prepare the file prior to pre-translation

Some recommendations should be considered when preparing a file that will be pre-translated using MT.

- **Remove unnecessary line breaks**: Sometimes sentences in a file are improperly truncated for various reasons, such as converting a PDF into an editable format or unintentional line breaks introduced during writing or DTP requirements.
- **Make the text format consistent**: To avoid a problem with the internal tags, the text format should be as simple as possible. A sentence with multiple tags is more likely to be incorrectly translated.
- **Strive for a balanced representation of terminology and domains or subdomains**: Consider the intended usage of the new model, such as a specific domain or subdomain it will be applied to. Aim to construct a representative dataset that encompasses an equal amount of data for each relevant option.

10. How to boost the MT power using a CAT tool as PECAT

Using a CAT tool to post-edit a MT empowers the process and helps to get much better results. Some of the improvements are the following:

- The CAT tool allows Translation Memories which can be combined with the MT in the pre-translation process.
- There is an improvement in consistency because CAT tools usually integrate QA check functionality.
- CAT tools file format (bilingual) can be re-used for further MT model training.