# Machine Translation Evaluation Project Based on an English-Chinese Parallel Corpus

Project integrity disclaimer:

*Mr. Wang Yun, "mogita", is a highly skilled full-stack developer based in Singapore and a long-time friend of the author. As the commit history suggests, mogita assisted in developing an ad hoc Python scraper and offered verbal guidance on file organisation methodologies. However, it is important to note that* ***mogita was not involved in any capacity in the corpora processing or the machine translation evaluation tasks.*** *He has never been asked or expected to do so, nor does he possess the expertise in these tasks.*

## Overview

This is a Machine Translation Evaluation project for the June 2023 Exam for the Machine Translation course at the University of Bologna.

The project starts with an ad hoc Python scraper based on Beautifulsoup to build **a parallel corpus of English and human-translated Simplified Chinese of political news from The New York Times Chinese site.** The date range is from 2020 to June 2023.

**A parallel corpus of 25k pairs of segments is obtained, with 1.1m English words and 2m Chinese characters**. Later, 12k pairs of segments are used as the training data, 0.3k pairs as tuning, and 0.5k as test data. ModernMT is used to train an adapted MT system to improve the MT translation quality. Finally, COMET, BERTScore, and BLEU are used for Automatic Evaluation, and a manual evaluation is also added.

Unfortunately, according to Automatic Evaluation results, the improvement can only be described as modest. A manual evaluation is also done, demonstrating some subtle improvement in the translation output of the Adapted MT, particularly in word order, thus resulting in a higher Fluency evaluation.

## Folder structure and file description

### Python scrapers

The Python scraper programme is situated in the root folder:

* links.py obtains links according to two regular expressions to obtain political news from The New York Times Chinese site.
* The obtained links are then put in links\_china.csv and links\_world.csv.
* articles.py scrapes the body text from individual pages.
* sample\_page\_unminified.html is an unminified sample page to aid project development.
* environment.yml provides a dependency list for Python environment configuration

### Corpora

Corpora are in the corpora folder. The initially obtained corpus is in two formats:

1. Original\_scrap\_output\_25k.csv
2. Original\_scrap\_output\_25k.xlsx

The original corpus is then divided into:

1. Corpus\_main\_12k.xlsxis also converted into 12k main-2358211.tmx translation memory exchange file and used as the Training Data.
2. Corpus\_Tuning\_and\_Test\_3.6k\_marked\_with\_colours.xlsx; all other sub-corpora with Test and Tuning in the filenames are derived from this file, whose names are self-explanatory.
3. Corpus\_unused\_10k.xlsxas a backup resource.

### Evaluation files

Evaluation files are in the evaluations folder.

* Four single-language .txt files from the Corpus\_Test\_0.5k sub-corpus, with the source, reference (human-translation), Baseline MT output, and Adapted MT output.
* Four Jupyter notebook files from two evaluation models.
* ibleu\_2023-06-16\_21-47-36 is the BLEU evaluation result output file.
* Manual Evaluation.xlsx is the 1-4 scale manual evaluation of two MT outputs.
* diffchecker-exported-pdf is a report file by the DiffChecker app to highlight the discrepancies between the Baseline and the Adapted text files.

## Automatic Evaluation Results

Only some modest improvements in the Adapted MT output were measured by the Automatic Evaluation methods.

* BERTScore model does not reveal any improvement nor deterioration in the Adatped MT output, as both Baseline and Adapted evaluated 0.873 against the Reference.
* COMET model indicates that the Adapted MT only improved microscopically, as it measures at 0.849 against Baseline's 0.848.
* BLEU even shows some deterioration, as the Adapted MT measured 5.86 against Baseline's 5.90.

*The author of this project would like to point out that, as a native speaker of Chinese, he disagrees with some of the scorings by BLEU on the individual segment level.*

## Manual Evaluation

The manual evaluation does give more credit to the Adapted MT output, but the improvement is still insignificant.

The manual evaluation is done on a 1 to 4 scoring scale to both outputs, on both Adequacy and Fluency dimensions. In all 519 segments, there are only 78 segments with discrepancies. The other 441 identical segment pairs are removed from the evaluation.

For Adequacy, the Adapted MT output scores 211 against the Baseline's 208; improvement is slightly more pronounced in the Fluency department, as Adapted MT scores 239 against Baseline's 226.

However, if all segment pairs are considered, and the default evaluation of the identical segments is 2.5, the Adapted MT scores in the Fluency dimension:

441 \* 2.5 + 239 = 1341.5

While the Baseline scores:

441 \* 2.5 + 226 = 1328.5

If we assume that the 1-4 scoring scale is linear, the Fluency improvement is about

1341.5 / 1328.5 - 1 ≈ 1\%

### Human Evaluation Bias

It has to be admitted that the author cannot avoid bias towards the Adapted MT output during the Manual Evaluation, as he is fully aware that this work will also be evaluated. This might contribute to a higher evaluation of the Adapted MT than it deserves.

————————

## Error Analysis in the Adapted MT Output

### Error Analysis

A significant part of the turning work is the reordering of words (in modern Chinese, a *word* is a unit of meaning that can comprise one or more Chinese characters), or in some cases, sub-sentences. This is clearly reflected in the Adapted MT output. However, the adapted ModernMT engine still **has difficulties correctly deciding when and on what part of the sentence to re-order, sometimes resulting in a worse translation than the Baseline.**

The Manual Evaluation shows that the Adapted MT output performs better in Fluency when the sentences are shorter. In other words, **the Baseline MT output suffers more from long sentences.**

## Conclusion

The author believes the ideal process to translate political news has already been implemented in some media organisations: Machine Translation, then post-editing by a person with journalism expertise and deep background knowledge in the reported matter.

Some mistakes by the Machine Translation in this project are made due to incorrect semantic understanding of certain verbs or nouns in the particular scenario, of which sometimes even a context-aware Machine Translation solution isn't capable.

The author thinks the advance of LLM might help in solving this problem. LLMs are trained with astronomical amounts of data and facts and can understand natural language and usages of particular vocabulary in specific scenarios. Of course, LLMs wouldn't be an immediate and perfect solution, as they themselves also hallucinate and provide inaccurate information. Still, they do reveal a promising future in further automating the translation process.