425 Van Buren St,
Monterey,
CA 93940

Pilot Project Proposal ↗

NMT training for UN Human Rights Council

Prepared by :

Océane Kristian
Kareem Gabriel
Xaneath

Prepared for :

the United Nations Office of the High Commissioner for Human Rights

April
2025

Updated Project Proposal

This pilot project sought to make an estimation of what would be necessary for commencing training of a neural machine translation (NMT) engine using SYSTRAN Model Studio Lite. This was done using data sets from within the Human Rights domain of the United Nations (French-English). Based on the predetermined criteria of the Post-Edited Machine Translation (PEMT) goals, the results are as follows:

Criteria	Goal	Actual Results
Timing	PEMT should be at least 40% faster than human translation.	PEMT is approximately 62.5% faster than human translation
Costs	PEMT should reduce translation costs by 30% compared to human translation while meeting the quality and timing goals defined above.	PEMT is approximately 16% cheaper than human translation
Quality	***Pass Threshold:*** *No critical errors. Errors in other categories must not surpass 7 errors per 500 words.* ***Fail Threshold:*** *One critical error.*	PEMTs have no critical error, but errors in other categories surpassed 7 per 500 words. ***Pass Threshold:*** *No critical errors. Errors of other categories should be within 10 with major errors not surpassing 2 errors per 500 words.*

April
2025

Actual Cost of Pilot Project

Task	Estimated Hours	Rate	Subtotal
Project Planning	20h	N/A	
Data alignment and cleaning	26h	$35	$910
Training, tuning and testing	43h	$45	$1935
MTPE	16h	$40	$640
Human evaluation	2h	$30	$60
Meetings	4h	N/A	
Total:			$3545

Instead of the initial estimated cost of $1415, the actual cost of this pilot program is $3545, more than two times of the estimated cost.

April
2025

Details of the Training Rounds:

Round #	Training time	Average score	Changes	Notes
1	5:42:55	N/A	Add data from 2020	
2	4:57:26	68.47	Add data from 2020, 2021,2022	
3	4:22:48	81.82 13.35	Add data from 2023	
4	4:09:02	81.65 -0.17	Realign data from 2020	
5	4:59:09	81.36 -0.29	Clean data from 2022	
6	4:48:30	80.75 -0.61	Clean & Realign data from 2023	
7	4:50:24	81.03 0.28	Clean & Realign data from 2020	
8	4:17:25	80.06 -0.97	Clean & Realign data from 2021	
9	4:33:52	81.51 1.45	Realign data from 2022	
10	N/A	N/A	N/A	

April
2025

Recommended Additional Training for Future Improvement

For the pilot project, our team spent 26 hours just aligning UN'S annual reports of the human rights council from 2020 to 204 in French with their English translations. Among them, the report of 2024 was carefully cleaned and processed as testing data; the rest were first roughly aligned as training data. During the 1st and 2nd rounds, we added all the data we had at hand; since the 3rd round, we intensively cleaned and realigned training data for fine-tuning. However, after the 3rd round, each of the following training turned out to have lower average score and average Systran score. After 43 hours of training, we stopped at the ninth round when all the improvement tries kept moving the score down. The human evaluation results from the first and last round also revealed a slightly declining quality of machine translation results, let alone the almost same amount of time spent on PEMT after each round of training.

We propose that the following steps may be considered to improve the system:

- Further break down the current paragraph-long segments into smaller units as sentences
- Except testing data, the rest data from 2024 should be added as training data
- The parent model should always be the model that has the highest score.
- Set up an auto QA checker to facilitate PEMT process
- Strictly document all the errors by categories and severity prescribed before
- The time spent on PEMT as well as the changes we made should all be documented closely
- There must be at least two linguists for PEMT and human evaluation tasks.

Currently, the data in our system is from the II. Resolutions and decisions brought to the attention of the General Assembly for its consideration and possible action in reports of 2020, 2021, 2022, 2023 and 2024. If we could have chosen a part with less long sentences in just one condition, we are likely to generate machine translations with less errors.

If the above mentioned steps were implemented, tripling the time spent on training, testing and tuning is the best-case scenario; which would largely increase the cost based on the actual cost listed above. Except for other projects we need to focus on, we need to work at least 3 hours in another 34 days.

Now we have 620 segments, each of which can be devided into at least 4 smaller segments; plus another around 800 smaller segments of 2024 data.

$$620\times4+800=3,280$$

The recent speed of alignment is 42 segments/h

$$3，280\div 42=6.667h$$

At least another 10 rounds of training are needed, thus the time for PEMT and human evaluation should be half of the sum of current

$$(16h+2h)\div2=9h$$

The average training time is around 5 hours so another ten rounds require

$$5\times10=50h$$

Plus at least 20 hours setting up a costumed QA checker and run for each round of training; meeting times(≥4h); document time(≥10h)

Total time would be 7h+9h+50h+20h+4h+10h=100h

If we must find some way to reduce the time we actually spent right now, first we need to use the same alignment tool so that the time each member spent on alignment should be close, not one person for 3 hours while another for 1.5 hours. We also need to have one more linguist for PEMT and human translations given the around 50,000 words of work.

Another biggest concern is that the metrics we could have (i.e. average score, average Systran score) is not solid enough because now when we rerun the same model Systran generated different scores, and duration of training to us.

April
2025

Observations and Recommendations

Based on the results from this pilot project, it is not recommended at this time to continue training this MT engine, and to instead continue on utilizing human translators for the needed translations. Factors in this recommendation included:

1. Incremental changes only in BLEU score

While we did see more of a significant increase from the second round of training and the third round of training, after that the change in BLEU score was infinitesimal, and actually decreased in several rounds. This indicates that the output of the MT engine was not actually improving despite our efforts. One way we mentioned to possibly improve the engine would be to utilize much more data, but it is not guaranteed to improve the BLEU score and MT output, while also posing its own challenges as mentioned next.

2. Time spent preparing training data

While there was plenty of data available in the form of reports and documents on the United Nations website, it was all in PDF format, requiring us to format and align the documents before it could be used as training data. This is very time consuming, and for the large amounts of data needed might be cost-prohibitive, given the amount of time needed for formatting and alignment.

3. Results of PEMT

While there was an improvement in the quality score between the first and last rounds of MT training (going from 24 points per 500 words to 14 points per 500 words), there was still a critical error found in the MT output from the last round, causing it to fail quality guidelines. It is estimated that there would need to be significant improvement in the MT engine before the output would be usable, even with PE. It was noted that there were many terms specific to the United Nations which the MT engine struggled with translating.

Also, the critical error found was in a sentence not being understandable, requiring a retranslation if it is to be understood. Thus rendering the output translation unusable for our purposes.

What We Would do Differently:

- Are NMT only improved incrementally over the rounds of training so while we were successful our process could be improved. In the future we could try to add more data in between each round.
- One thing we learned early on is that we need to make sure there is enough testing data. The first batch we used was too small.
- The UN website was perfect for getting data. They have publicly available documents in multiple languages. It made it easy to select and align documents.
- It might be beneficial to have more than one member of our team be a speaker of the source language. With only one French speaker there was potential for bottlenecks in the process.

April
2025

Final Delivery

01

1. Updated proposal on trained NMT for UN annual reports of the Human Rights Council

02

2. Details of the data used for pilot training (training, tuning, and testing data sets)

03

3. Iteration records for training, tuning and testing and training scores

04

4. A review of this training with improvement suggestions

April
2025

Terms and conditions

- The pilot project is intended for evaluation purposes only

- Both parties agree to maintain strict confidentiality of all information, communication content, and internal processes shared during the pilot.

- Participation in this pilot does not create any obligation for the client to enter into a long-term service agreement.

Client Signature & Date:	
Representative, Human Rights NMT project Signature & Date:	