Machine Training Project Proposal

FOR: Kyoto Tourism Federation
〒602-8570, Kyoto Prefectural Office,
Yabunouchi-cho, Shinmachi, Shimodachiuri,

Good Vibes Tours
460 Pierce St,
Monterey, CA 93940

Olivia Plowman, Albert Phan, Naomi Stock, Aleria Amaral

Summary of Pilot Project Outcomes

For our MT project, we started training a machine that we hoped would make PEMT 40% faster and cheaper than human translation, with the project quality being within acceptable score range as defined by the Multidimensional Quality Metrics Framework. Over the course of our two week training process, we completed a total of 12 rounds using a selection of 45 different instructions for use of tuning, testing and training. The initial round of training achieved a BLEU score of 18.24, which increased to 19.05 in our best round, a difference of 0.81 points. We attempted different tactics during each round to improve the quality of our NMT engine.

We initially struggled to put our XML into Custom Translator as we were trying to get rid of all the tags through regex. While this did take roughly 5 days of cleaning around 14 thousand documents, we were able to effectively proceed with Custom Translator. Being able to pick out the data we needed would have been helpful, however. We cleaned all of it with the mindset that we would use its entirety. After initial cleanup, we proceeded further by deleting extra segments. We did this step through Notepad++ by simply deleting extra lines. Probably the most difficult step was working through memoQ with segmentation, as we separated some segments and deleted others to make sure aligning would proceed without issue.

We concluded that tuning does not make much of a difference in terms of size, but as testing increases it will lower the BLEU score. The more clean training data we used the better it performed.

Proposed Objectives

Our objectives for this project were to train a neural machine translation for English-speaking tourism in Japan to provide Japanese to English translations. We wanted our post-edited machine translation to meet the following criteria, listed below:

- Efficiency: PEMT 40% faster than human translation
- Cost: PEMT 40% savings over human translation
- Quality: PEMT with an acceptable score based on the MQM Framework

Efficiency

Our baseline to compare against is that humans can translate at roughly 400 words per hour. For each quality check we conducted, we took roughly 1000 characters (including spaces) from the testing data and timed how long it took to identify errors and post-edit the translation. Quite miraculously, they each rounded out to about 15 minutes each.

- QA1: 184 words/15mins, or 736 words per hour
- QA2: 178 words/15mins, or 712 words per hour
- QA3: 198 words/15mins, or 792 words per hour

These times give an average of about **747 words per hour.**

This results in about an **86% increase in efficiency** compared to straight human translation. This is over twice the amount of efficiency that we initially presumed.

Cost Savings

We are assuming that a human translator would cost $.20/word for translation and $50.00/hour for PEMT. If this is the case, when using our trained NMT engine, it would cost $.05/word (40% less than for HT) for translation and remain at $50.00/hour for human PEMT.

For the same amount of words (e.g. 180 words), it takes a human roughly 30 minutes to translate at a rate of 400 words per hour. If we assume that post-editing takes approximately ⅓ of the time it takes to translate, the post-editing would take 10 minutes.

Human Translation (HT): $.20 x 180 words = $36.00
HT Post-editing: $50.00 x .17 hours = $8.33
Machine Translation (MT): $.05 x 180 words = $9.00
PEMT: $50.00 x .17 hours = $8.33

TOTAL = $44.33 for HT
TOTAL = $17.33 for MT

This means that there is a **total cost savings of 39%**. This is 1% below our goal of 40% savings, however we believe that to be negligible.

Quality

Our quality estimations were formed by creating a variation of the Multidimensional Quality Metrics (MQM) Framework that tested for accuracy and fluency in two sets of roughly 500 characters each taken from the target testing data. We set them on a 0-100% scale, with 82% being a passing score. Errors were assigned points based on severity:

Minor Errors: Technically errors but don’t disrupt the flow or hinder comprehension. (1- 2 points)

Major Errors: Disrupt the flow, but what the text is trying to say is still understandable. (5 points)

Critical Errors: Inhibits comprehension of the text. (9 points)

Here is our set of examples of errors through MQM Framework:

Category	Minor Example	Major Example	Critical Example
Accuracy			
Mistranslation	/	Wrong pronouns	Still in Japanese
Untranslated	/	Includes romanizations	Still in Japanese
Omission	/	Awkward omission	Unintelligible omission
Addition	Added small amount of in-context words	Added out-of-context words	Added Japanese
Fluency			
Unintelligible	/	/	Example: ‘fklsajdf;’
Mechanical			
Spelling	Missing accent marks/inconsistency in romanization spellings	Distracting misspelling. Ex. “wavesss”	Misspelling makes word unintelligible
Typography	Extra spaces	Improper typography Ex. No question mark	Japanese punctuation
Grammar	Still able to discern the meaning of the sentence Ex. “I” becoming “we”	Not following grammar conventions, such as run-on sentences	Unintelligible sentence due to jumbled grammatical errors
Locale Convention	Spelling conventions incorrect	/	Critical misuse of terminology from different locale that could lead to misunderstanding

Our first check of quality assurance was performed after the second round and was based off of Model 2 of the training rounds. Even though one of the two sets of about 500 characters contained a critical error, the scores were **96%** and **91%**, a passing score.

Our second check of quality assurance was performed after all rounds of training were finished and was based off of our last round, Model 12. The scores were **66%** and **75%**, which both fail.

Our last quality assurance check was based on our highest-performing round, Model 5. The scores were **97%** and **91%**, which pass with some improvement over our first QA check.

In essence, our NMT at its best passes quality checks.

Details of Training Rounds

Models	Date	Bleu Score	*Training Data/File Type	*Tuning Data/File Type	*Testing Data/File Type	Comments
Model 1	3/18/21	18.24 ⇨	8,279 seg./txt	435 seg./tmx	415 seg./tmx	Training data was two separate txt files
Model 2	3/21/21	17.15 ⇩	8,279 seg./txt	546 seg./tmx	586 seg./tmx	Except increase of ~100 tuning/testing segments, unchanged from Model 1
Model 3	3/23/21	16.8 ⇩	8,708 seg./txt	619 seg./tmx	654 seg./tmx	Increase of training data, ~100 tuning/testing segments
Model 4	3/23/21	18.07 ⇩	8,708 seg./txt	435 seg./tmx	415 seg./tmx	Reduction to of tuning/testing to see if it would help the score
Model 5	3/25/21	19.05 ⇧	6,211 seg./tmx	435 seg./tmx	415 seg./tmx	Change of training data to tmx bitext, tuning/testing unchanged
Model 6	3/25/21	18.63 ⇧	6,211 seg./tmx	546 seg./tmx	538 seg./tmx	Training unchanged, ~100 tuning/testing increase
Model 7	3/27/21	19.05 ⇧	6,211 seg./tmx	569 seg./tmx	415 seg./tmx	Reduction of testing data to see if it would improve score
Model 8	3/27/21	18.8 ⇧	5,400 seg./tmx	435 seg./tmx	415 seg./tmx	Used cleaner files for training data, tuning/testing back to baseline
Model 9	3/27/21	15.28 ⇩	5,901 seg./tmx	310 seg./tmx	415 seg./tmx	0 tuning files selected as experiment
Model 10	3/27/21	11.1 ⇩	5,901 seg./tmx	437 seg./tmx	310 seg./tmx	0 testing files selected as experiment
Model 11	3/27/21	18.62 ⇧	6,211 seg./tmx	619 seg./tmx	461 seg./tmx	Baseline training/testing data, increase of ~200 segments in tuning
Model 12	3/27/21	18.84 ⇧	6,032 seg./tmx	435 seg./tmx	415 seg./tmx	Increase of cleaner training data, =baseline tuning/testing

*Note: Segment count used by MS Custom Translator

In our initial projections, we planned to start training with 12,000 sentences and increase over time to 16,000 sentences. This proved to be difficult to achieve, as the segmentation from our corpus did not match up with the projection of 12,000 or 16,000 sentences. Additionally, the limitations given to us by Microsoft Custom Translator for having a minimum of 10,000 segments and a maximum of 2 million characters per Model constrained our ability to experiment with the amount of segments used for training. When we increased the segments too far beyond 10,000, we ran into the character limit. For the tuning and testing data sets, while we did have enough aligned documents for 4000 sentences minimum and 8000 sentences maximum, we found over the course of our testing that increasing the tuning and testing data in the form of aligned tmx files did not seem to improve our BLEU scores. As a result, we started with a minimum of around 500 segments each for tuning and testing and a maximum of around 800 segments, well below even our projected minimum.

One of the things that we learned over the course of our project was that MS Custom Translator does not use all of the segments we attempt to utilize for a round. This means that even in our best round, Model 5, although we tried to use 10,582 pre-aligned segments, after MS Custom Translator re-aligned them, only 6,211 segments were used in training. There was a similar phenomenon among tuning and testing segments where each time approximately 100 less segments were used for training than we attempted to use.

Actual Timeline and Costs

In our original timeline, we wanted to begin with a start date for our training rounds on March 2nd, but that was greatly delayed due to the unexpected amount of required data clean-up. Corpus data needed to go through several iterations of clean-up and conversion before they could even be uploaded to Custom Translator. This became our bottleneck in training efficiency.

Of the 14.1k corpus data available to us, we could only use a fraction of the files in order to stay below the 2 million character limit for training data in the pilot setting. Furthermore, of the 15 hours allocated for document alignment and clean-up, all 15 hours were used only for clean-up, and as such additional hours would need to be added for aligning tuning and testing data. MT training round time was also underestimated, with each round taking roughly 1 to 1.5 hours. A service outage in the middle of one of our training days also forced us to push subsequent rounds into our Post-Editing and QA days.

Task	Estimated Hours/Quantity	Quantity	Total # of Hours	Rate	Subtotal
Document Alignment & Clean-up	5	5	25 hours	$50.00	$1,250
MT Training Rounds	1	12	12 hours	$50.00	$600
Glossary Creation	4	1	4 hours	$50.00	$200
Post-editing	1.5	3	4.5 hours	$50.00	$225
Human Evaluation/QA	1.5	3	4.5 hours	$50.00	$225
Project Management	**Flat fee**	1	--	15% of Subtotal	+$375
				TOTAL	$2,875

For a full project, the amount of data required would be higher, and thus the 15 hours spent on clean-up would also be proportionately increased as well. Just doubling the data count would equate to roughly three or more days of clean-up for a full-time employee, but we would still keep our rate of $20 per hour before markup (to reach the post-markup rate of $50 in our pilot proposal).

Recommendations for Continued Training

We recommend that continued training be based off of Model 5, which was the highest-performing training data set.

We believe that what made this set perform so well was the fact that we changed our training data to tmx bitext while leaving our tuning and testing data unchanged. Therefore, this step was integral to our efficiency and effectiveness, and we would recommend using more varieties of file types in order to have a wide variation of tuning, testing files included into one training round. We would recommend trying different variations of MS Custom Translator which would not be limited to a certain number of characters per round. It would also be advisable to take advantage of a wide variety of file types (docx, txt, tmx, etc.). We also strongly recommend using more duplicate segments, perhaps even in different file types in the same round to help improve the BLEU score.

For example, we suggest doubling the amount of segments in the training data set that we used (as this seems to make the most difference in the quality of the testing translation). Our original estimated workflow is as follows:

$$\frac{6{,}211(training) + 415(testing) + 435(tuning)\ segments}{17\ hours} = 415 \text{ segments per hour}$$

The original sentence count of 10,582 (decreased to 6,211, as this is the number used by Custom Translator) would come out to 12,422 used segments total in this case. This would further increase the accuracy of our engine while being able to produce a higher number of segments per hour.

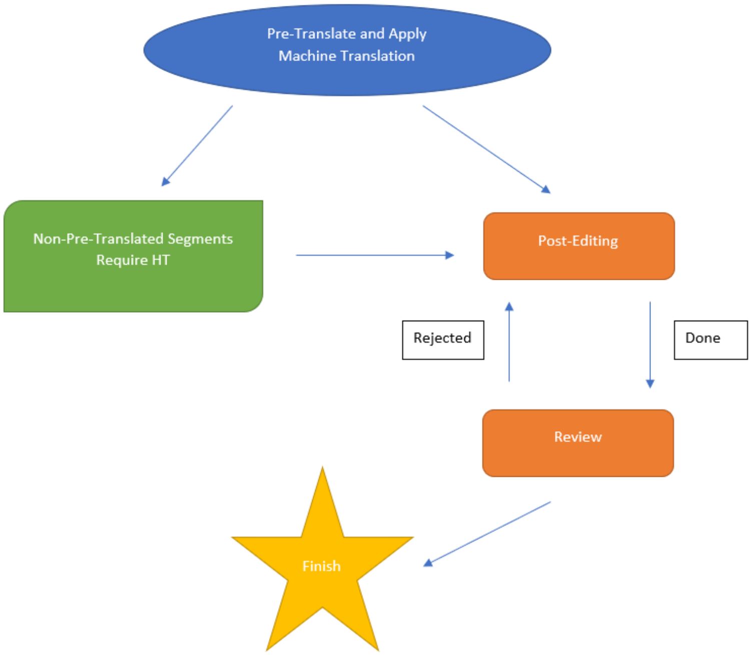
$$\frac{12{,}422(training) + 415(testing) + 435(tuning)\ segments}{20\ hours} = 664 \text{ segments per hour}$$

We should also consider improving the quality of the pre-aligned segments so that when we utilize them as training data in a round, MS Custom Translator ends up using more of the segments for training.

How You Might Use this NMT in a CAT Tool

This NMT is intended to be used, as the name suggests, primarily as a machine translation tool. Our NMT has been trained to be used for Japanese to English translation, primarily in the tourism sector. We suggest that you use this NMT when translating along those parameters.

Ideally, this NMT would pre-translate segments that it believes to have about an 82% fuzzy match (to align with our MQM Framework for quality), and provide only suggestions for segments with a lower match rate so the translator may finalize the initial translation and post-edit. Once finished, the post-editing would be sent to a reviewer, and either accepted or rejected for further post-editing.


Pre-Translate and Apply Machine Translation
Non-Pre-Translated Segments Require HT
Post-Editing
Rejected
Done
Review
Finish

The NMT would also include automatic QA checks to assist PEMT/QA processing. This way, the translator and/or post-editor would receive warnings or errors if punctuation were missing or different (especially if it were accidentally Japanese punctuation),etc.

Anticipated Results

Were this NMT able to be fully trained, we predict that it would mostly put out machine translations that pass our quality assurance tests with at least an 82%. Of course, no machine translation can be perfect, and based on our three QA checks, this NMT produces sentences that pass muster about two-thirds of the time. It may occasionally put out unintelligible sentences that require a human translator to decode from the source Japanese. However, given that the efficiency of PEMT with this NMT is so high, we expect efficiency to remain high even with the occasional hiccup.