Machine Training Pilot Project Proposal

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

Good Vibes Tours
460 Pierce St,
Monterey, CA 93940

Our Objectives

With this pilot proposal project, Good Vibes Tours aims to train a statistical machine translation (SMT) for English-speaking tourism in Japan. Our goal is to provide Japanese to English translations, and our post-edited machine translation (PEMT) from this translation will meet the following criteria:

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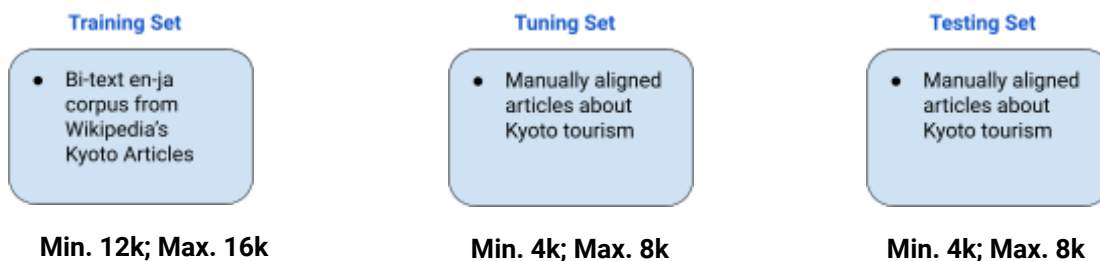- Quality: PEMT with an acceptable score based on the MQM Framework

Post-editing will be done by a human reviewer using the **Multidimensional Quality Metrics** (MQM) Framework for **accuracy and fluency** set on a 0-100% scale. Each error will be assigned a score of 1 or 2 depending on severity of minor and major respectively, and subsequently subtracted from the initial score of 100%. Critical errors are 9 points. In order to pass a quality check, the score must be 82% or higher. The equation is as follows:

Translation Quality (TQ) = 100 - Accuracy (AP) - (Fluency of Target(FPT) - Fluency of Source(FPS))

The engine will be trained over the course of three weeks. Upon completion of this training process, estimates pertaining to costs and requirements for a full MT training project will be made upon assessment of rate of improvement and achieved levels of quality.

Data Sets

We will be splitting our data sets into three categories: training, tuning, and testing. The training sets will have the largest data volume stemming from pre-aligned bi-text documents. The tuning and testing sets will be made up of manually aligned bi-text documents. Each model will have a limited number of sentences, coming to a maximum of **16,000 sentences** for training, **8,000 sentences** for tuning, and **8,000 sentences** for testing. We plan to start our model training with **12,000 sentences** in the training set and **4,000 sentences** in tuning and testing respectively.

Training Set
- Bi-text en-ja corpus from Wikipedia's Kyoto Articles

Min. 12k; Max. 16k

Tuning Set
- Manually aligned articles about Kyoto tourism

Min. 4k; Max. 8k

Testing Set
- Manually aligned articles about Kyoto tourism

Min. 4k; Max. 8k

Process

Working from a selection of 14.1K [bilingual corpora of Wikipedia's Kyoto Articles](#) in XML file format, the engineers will begin a first round of MT training using 10 bi-text documents for training, 1 for tuning and 1 for testing. This plan assumes an average of about 14,000 sentences per document. The MT system will be trained repetitively over the course of successive rounds, with different tactics administered to improve the system's BLEU score. These strategies may include increasing the number of training and tuning documents, adding target-only data and adding a dictionary. We will record a BLEU score every round, and will conduct over 12 rounds (including rejected rounds).

After all of these rounds have been completed, estimates regarding cost, quality and time will be reassessed. The MTs will be downloaded and post-edited by four reviewers for 30 minutes each. The number of words that are able to be edited during this time will be compared to the standard 300 words/hour for human translation (HT). Cost savings will also be evaluated based on a standard $0.25/word rate for HT and a $0.15/word rate for PEMT. The two translations will then be exchanged and we will assess their quality using a version of MQM QA Model.

Timeline

4 Training Rounds	Post-Editing and QA	4 Training Rounds	Post-Editing and QA	4 Training Rounds	Post-Editing and QA	Savings Estimate	Lessons Learned
March 2nd-9th	March 10th	March 11th-18th	March 19th	March 20th-25th	March 26th	March 27th	March 28th

Costs

Task	Estimated Hours/Quantity	Quantity	Total # of Hours	Rate	Subtotal
Document Alignment/ Cleaning	5	3	15 hours	$50.00	$750
MT Training Rounds	0.5	12	6 hours	$50.00	$300
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	+$225
				TOTAL	$1925

Deliverables

Upon completion of the pilot project, the following will be delivered to the client:
- Data gathered from data sets used for training the engine (including but not limited to documents used, changes made and the BLEU scores for each round)
- An updated proposal using the data collected from post-editing and QA that will evaluate the achievability goals outlined in the pilot proposal
- Recommendations for ways to continue training the MT engine as well as a proposed workflow form utilizing the engine for use with a translation management system