

SMT Training Updated Project Proposal

S.H.E.L10N

Frances Chang | Clinton Lin | Jenny Lowe | Charlene Wang | Valerie Yin

OBJECTIVES

As stated in the pilot project proposal, this project is intended to train a statistical machine translation (SMT) engine in Microsoft Translator Hub to translate the Taiwanese Laws from Traditional Chinese into English. A total of 6 rounds of successful training were completed with an initial data set of approximately 10,000 segments used as training data, approximately 2,200 segments used as tuning data, and approximately 2,000 segments used as testing data. The initial round of training achieved a BLEU score of 6.19 and over the next two weeks, 12 more rounds of training were done with a result of 32.41 (~70%) as the best BLEU score achieved.

SYSTEMS	TRAINING	TUNING	TESTING	BLEU SCORE
#1	1	1	1	6.19
#2	1	4	4	27.63
#3	2	3	6	FAILED
#4	2	3	6	6.28
#5	1	4	7	FAILED
#6	4	4	4	FAILED
#7	4	3	5	8.33
#8	10	4	4	32.21
#9	13	4	1	32.41

In the pilot proposal, we estimated that post-edited machine translations (PEMT) would be 50% cheaper and more efficient than human translation (HT). However, as the following tables indicated that those goals were close, but not met. Thus, the goals have been lowered accordingly with an updated percentage of 35% and 40% respectively.

COST

TASK	RATE (PER WORD)	SUBTOTAL
HT	\$0.12	\$60
PEMT	\$0.06	\$30

REVIEW	\$0.05	\$25
HT+REVIEW	-	\$85
PEMT+REVIEW	-	\$55
COST VARIANCE	\$30	
COST SAVED	35.3%	

PRODUCTIVITY

SAMPLE OF 500 WORDS	POST-EDITOR FOR MT	HUMAN TRANSLATOR
PEMT	1 hour	-
HT	-	3.5 hours
REVIEW	1 hour	1 hour
PEMT + REVIEW	2 hours	-
HT + REVIEW	-	4.5 hours
TIME VARIANCE	2.5 hours	
TIME SAVED	41.7%	

QUALITY

In the pilot proposal, we stated that the Localization Industry Standards Association (LISA) QA Metric would be used to evaluate the PEMT and HT, however, we determined that the Quality Insurances (QI) Metrics from WeastO, a Taiwanese translation agency specialized in legal and IT documents would be a more appropriate metrics for this project. The QI Rate must not exceed 10 to be considered a satisfactory score. Two reviewers evaluated the PEMT with a average result of 19 from a score of 22.68 and 15.12 respectively. The result could be seen from the following two tables.

Although neither the average score nor any individual score were satisfactory, we felt that we came pretty close with an engine that has only a BLEU score of 32. We feel that with a better BLEU score, we can meet and surpass the goals that we set.

REVIEWR #1			
DATE: 4/24/2017		TOTAL WORDS: 529	
ERROR CATEGORY			
ERROR CATEGORY	ERROR WEIGHT	COUNT	SUBTOTAL
Terminology	1	4	4
Mistranslation	1	2	4
Accuracy	1	4	4
Consistency	1	0	0
Style	0.5	0	0
Language	0.5	4	2
		TOTAL	12
$\text{QI Rate} = \frac{\text{Total points of error weight}}{\text{Total word count}} * 1000 \text{ (per mill)}$			
QI RATE	22.68		

REVIEWER #2			
DATE: 4/24/2017		TOTAL WORDS: 529	
ERROR CATEGORY			
ERROR CATEGORY	ERROR WEIGHT	COUNT	SUBTOTAL
Terminology	1	4	4
Mistranslation	1	0	0
Accuracy	1	1	1
Consistency	1	0	0
Style	0.5	3	1.5
Language	0.5	3	1.5
		TOTAL	8
$\text{QI Rate} = \frac{\text{Total points of error weight}}{\text{Total word count}} * 1000 \text{ (per mill)}$			
QI RATE	15.12		

RECOMMENDED ADDITIONAL TRAINING

Although none of our initial goals were met, the pilot project has taught us many lessons for future SMT engine training.

- 1) To continue the training, well-aligned source documents will be needed either from the client directly or downloaded from trusted sites. In our initial training, the documents had line breaks in between sentences, which resulted in very badly aligned parallel texts. The documents used in the pilot project were mostly .xlsx files, however, later in the training process, monolingual texts in PDF format were used with favorable results.
- 2) For training static law documents, adding a dictionary would likely be beneficial to the engine. Many repetitive terms were used throughout the documents as well as organizations and mentions of other laws. Adding those terms in the dictionary could only improve the output.
- 3) Use of Computer Assisted Translation (CAT) tools for document alignment.
- 4) During the pilot project training, we noticed that the BLEU score improved significantly between the first and second round of training when documents with better aligned (hence better quality) were added to the Testing set. The BLEU score increased significantly again between the seventh and eighth round of training when a large amount of documents, regardless of quality, were added to the the Training set. Thus, we concluded that in the Testing set, quality > quantity, while in the Training set, the opposite is true.

TIMELINE AND COST

A minimum of 100,000 segments are suggested in the Training set to fully train a SMT engine, however, with this specific topic in mind, at least 500,000 segments should be used to achieve a more realistic result. Due to a lack of good quality source documents available, an exact timeline would be difficult to established. Based on the amount of segments we used in the pilot project, a fully trained SMT engine would take at least a 3 weeks time to properly find, prepare and train 100,000 segments with a full-time employee dedicated to this task. An estimated cost is shown in the table below.

TASK	HOURS	RATE	SUBTOTAL
DOCUMENT PREPARATION AND MT TRAINING	120	\$35	\$4,200
PM FEE	10%		
HT	\$0.12/word		
PEMT	\$0.06/word		
REVIEW	\$0.05/word		
TOTAL			\$4,620+

ANTICIPATED RESULTS

Based on the result of the pilot project, a fully trained SMT engine is expected to yield equivalent or better results than the initial pilot objective goals in terms of productivity, quality and cost. With continuous training with better quality documents and a vast dictionary, the BLEU score is expected to continue to improve, thus resulting in better quality output. The engine would also be expected to deliver translated material that should only take a limited amount of time to post-edit. The pilot project proved that a fully trained SMT engine would save at least 35% of the total costs compare to that of a human translator and review. The results also proved that the engine would be 40% more productive than a human translator.

Signatures



(VENDOR - S.H.E.L10N)



(CLIENT - Adam Wooten)