KOT Translations

—

Training an SMT Engine

Project Overview

- Train an SMT engine to translate press releases for website of US Embassy in Moscow
- Perform ten rounds of training and a post-edit/QA evaluation

Client: State Department/US Embassy Moscow

Domain: International relations (general audience/informational)

Pilot Project Goals

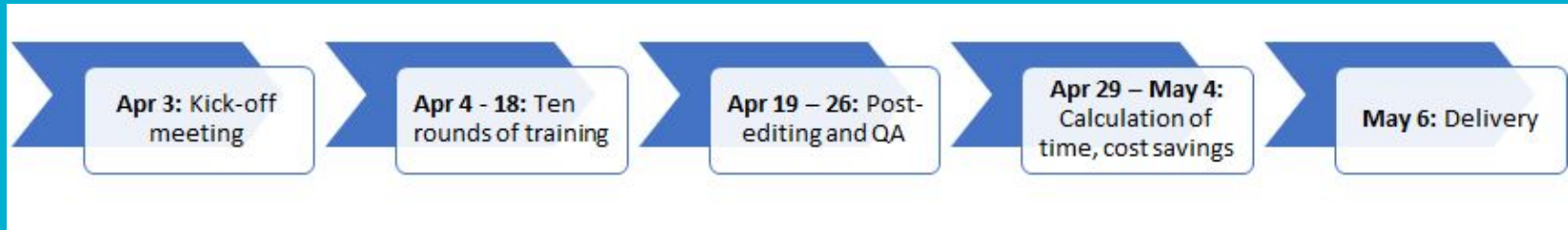| Efficiency	Cost	Quality
20% efficiency	20% savings	High enough quality to not cause an international incident--below 10 points with no critical errors.

Updated Objective and Goals

- Retrain the engine by removing non-State Department IP (bitexts from UN) and adding bitexts provided by the State Department related to US/Russia relations.

Efficiency	Cost	Quality
66% efficiency	44% savings	High enough quality to not cause an international incident--below 10 points with no critical errors.

Updated Data Sets

Glossary	Training Data	Tuning data	Testing data
Glossary file with names of people key to U.S.-Russia relations and abbreviations and full names of relevant organizations, agreements, etc.	A large volume of bitexts (at least 10,000 sentences) received from the State Department of texts related to U.S.-Russia relations	Bitexts of various kinds of texts filed under "press releases" on the U.S. Embassy's website	A compilation bitext file with all the different kinds of documents filed under "press releases" on the U.S. Embassy in Moscow's website

Timeline

| Apr 3: Kick-off meeting | Apr 4 - 18: Ten rounds of training | Apr 19 – 26: Post-editing and QA | Apr 29 – May 4: Calculation of time, cost savings | May 6: Delivery |

Training: Process

File Prep	We have files for testing, tuning, and training already aligned, and we are requesting an additional bitext file or English-Russian file pair of at least 10,000 sentences from the U.S. Department of State so that we can remove the outside training data (U.N. bitext files) from our training.
Training	BLEU score will be evaluated each round. Tweaks will be made to the training and tuning rounds: additional texts, aligning some of the training data.
Assessment	Assessments will be made by Russian lead linguist, who will also do the post-editing. The metric above will be used to assess quality. Post-editing time will be compared to average speed of translation per page--30 minutes for a high-quality translation. Cost savings will be determined using the metric of $45/hr.

Round	Change	BLEU Score Baseline: 28.25 +2.26
1	-	30.51
2	add more training from UN bitext file (aligned file of 2,992 sentences)	29.9
3	use model 1 and add tuning file (aligned file of 133 sentence pairs)	30.3
4	use model 1 and use tuning file from model 2 as training	29.7
5	use model 1 and add training file of 51 new aligned sentences from the newest texts on website	30.05
6	use model 1 and add glossary with names and abbreviations	30.3
7	use model 1 and add texts from both model 3 and model 5 as training	29.95
8	use model 1 and align one of the training texts	30.29
9	use model 1 and align another one of the training texts	30.24
10	use model 1 and align two of the training texts	29.67

Training: Lessons Learned

- You need a lot of texts to make a real difference in the BLEU score
- New limitations of Microsoft Custom Translator: 10,000 sentences, even if you align 2,000+ plus
- Limitations of BLEU score: model with glossary was slightly lower, but post-edit time was greatly reduced, so we are recommending implementing one, despite the slightly lower BLEU score

Post-Edit Evaluation: QA Metric

	Minor	Major	Critical
Language	1	5	10
Accuracy	1	5	10
Style	1	5	10
Terminology	1	5	10

Adapted from *TAUS Harmonized DQF-MQM Error Typology*:
https://www.taus.net/evaluate/qt21-project?fbclid=IwAR3rDS8
2ah01Ba7KT3ysU7BKnpn4PvhAU5lqhurd4SzqnsEyjYsAFnLw
L0Q#harmonized-error-typology

Post-Edit Evaluation

QA	Model 1 (BS 30.51)	Model 6 (BS 30.3)
Language	4	3
Accuracy	4	3
Style	1	1
Terminology	1	1
Total:	10	8
Post-edit estimate	2 hours	1.5 hours

Estimated Costs for First Year		
Month	**SMT Cost**	**HT Cost**
Month 1	$1745.50	$405
Month 2	$90	$405
Month 3	$90	$405
Month 4	$90	$405
Month 5	$90	$405
Month 6	$90	$405
Month 7	$90	$405
Month 8	$90	$405
Month 9	$90	$405
Month 10	$90	$405
Month 11	$90	$405
Month 12	$90	$405
Total:	$2713	$4860
TOTAL COST SAVINGS: 44%		

Post-Edit: Lessons Learned

- Results were very good.
- No critical errors and under ten points in QA metric
- Greater efficiency and cost savings compared to human translation
- Glossary helps with post-edit process

SMT Training Cost Breakdown: Estimates

Task	Estimated Hours	Rate	Subtotal
Document alignment	15	$45.00	$675.00
MT training	15		$675.00
Glossary creation	10		$450.00
Post-editing	4		$180.00
QA	4		$180.00
			Total: $2160.00

SMT Training Cost Breakdown

Task	Estimated Hours	Rate	Subtotal
Document alignment	22.5	$45.00	$1,012.50
MT training	3.3		$148.00
Glossary creation	10		$450.00
Post-edit evaluation	2		$90.00
QA	1		$45.0
			Total: $1745.50

Deliverables

- SMT engine ready for deploy
- All associated files, including glossary, so engine can be updated and maintained
- Records of results and BLEU score and post-edit evaluation

Key Lessons Learned

- BLEU score does not tell the whole story
- Glossaries have their place
- Domain and text type play a big role in SMT engine success
- You need a lot of files and a lot of texts to add to engine