BESST Translations
460 Pierce St
Monterey, CA

SMT Training Project Proposal

ATTN: Adam Wooten, Founder & CEO of Wooten Inc.

SUMMARY OF PILOT OUTCOMES

We launched the three-week pilot project following the kickoff meeting on March 30, 2016. Over the course of three weeks, we trained an MT engine for the purpose of translating KDDI financial statements from Japanese to English. On March 31, we constructed an initial translation engine and perform an initial test. Initially, we trained a system using bilingual pairs of PDF files, totalling approximately 10,000 segments, that came from financial statements of Softbank, a rival company. We tuned the MT engine using bilingual pairs of PDF files, totalling approximately 3,000 segments, that came from recent KDDI financial statements. The most recent KDDI year-end financial statement was used to test the MT engine. Upon receiving an initial BLEU score of 14.36, we performed 9 subsequent training sessions, resulting in a final BLEU score of 22.03.

Through the subsequent 9 training sessions, we found improvements in the MT engine through the following actions: converting files from PDF to HTML format, manually aligning the documents, adding monolingual documents (using year-end statements from Verizon, AT&T, T-Mobile, and Sprint), and cleaning the bilingual data, for example by removing extraneous spaces. One method that did not appear to improve the MT engine was removing numbers; however, this method warrants further exploration during the project, as an examination of the output has shown that numbers are one of the factors resulting in poor results in certain places.

We also ran the pilot machine translations through human post-editing, giving two editors the translations to edit, and recorded the time it took them. The results are listed in the “Project Objectives” section. As the goal was to achieve a usable translation of a reasonable quality, while saving time and costs, we feel that the post-editing results meet our criteria to move forward with the project.

PROJECT OBJECTIVES

As described in our initial pilot project proposal, we had three objectives that we were hoping to accomplish during our initial training phase.

- **Quality goals:** A machine translated financial statement that is not only readable but also does not require an excessive amount of human post-editing or QA. We will grade the quality of the MT output using the American Translators Association (ATA) certification error guidelines because of their clearly defined scoring metrics (see chart on next page). The MT output should have no 16-point errors, and overall no more than 20 error points total per 250 words of target text.
- **Efficiency goals:** Achieve a machine translation that, after human post editing, is significantly faster than an entire human TEP process. Specifically, it should be 70% faster than human translation.
- **Pricing goals:** Our machine translation should result in a significant price reduction in translation of Japanese financial statements to English. We expect a 70% reduction in cost compared to human translation.

In order to take a look at how far along we are in achieving these goals we need to take a look at the efficiency of our PEMT text and decide if the speed and price reductions are feasible. The following table helps to show the data that we collected from our team's PEMT using the machine translation engine that we trained during the pilot phase.

	Post-editor #1	Post-editor #2	Average
PEMT words/hour	1400	1600	1500
PEMT Time needed for 2,000 words	1 hr 25 min	1 hr 15 min	1 hour 20 min

For comparison, the time required for human translation is shown in the table below:

	Average
HQ Translation - words/hour	250 words/hr
HQ Translation - time needed for 2,000 words	8 hours
HQ Translation - words reviewed per hour	2000 words/hr
HQ Translation - time to review 2,000 words	1 hour
HQ Translation - Total time to do 2,000 words	9 hours

Our efficiency is thus calculated to be 85%. This is still only an initial estimate but it does fall within our efficiency goals as outlined above.

For a breakdown of our pricing objectives we can look at the following table:

	Rate per word for TEP	Estimated cost per 2,000 words
High Quality	$0.30 per word	$600
PEMT	$0.10 per word	$200
	PEMT Savings:	$400 ~67% savings

Thus, our goal of achieving a nearly 70% reduction in cost is close to being a reality. However, this only takes into account the pricing that we would be offering the client after we have sufficiently trained our machine translation engine, a process that does cost a significant amount of time and money. As a result the actual savings will differ based on the amount of words needed to translate in order to "break-even" when factoring in the costs for training the machine translation engine.

RECOMMENDATIONS FOR CONTINUED TRAINING

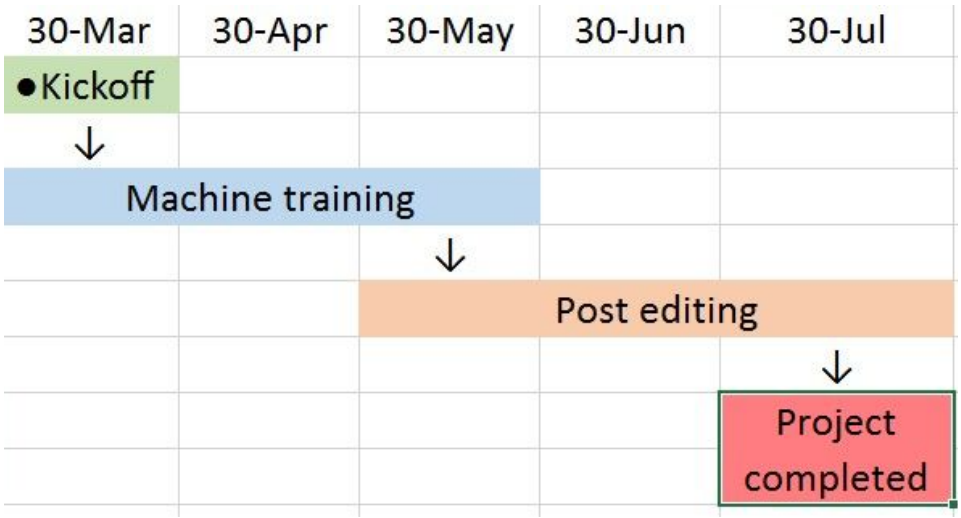
Our research into training a machine translation engine may not have produced a very high initial BLEU score, but it has given us a wealth of information to draw upon in our future endeavors.

1) If we were to continue this process our group believes that contacting the client, KDDI, directly and asking for source documents would be of great benefit. One of the issues that we ran across was poor segmentation caused by using PDFs that were publicly available rather than much cleaner source files.
2) Another recommendation our group had was adding more documents to our training and tuning set. During the pilot, we worked with approximately 10,000 segments, meaning the minimum required. However, a more realistic minimum would be closer to 100,000 segments, and in this project, we would hope to obtain a number closer to 1,000,000. When training a machine translation engine, having more is often times better because it will give the machine more possible matches to use when trying to translate the test documents. The key with this suggestion is to pick relevant documents that will add something to our machine engine, and not just adding documents for the sake of adding them.

3) Alignment. Many of the documents that we used were poorly aligned, but considering this was a pilot project we did not have the time our resources available to align every single document that we were using. If the project were to move forward, spending time in alignment would likely yield good results.
4) Lastly, our group had some small success adding monolingual texts to the pilot machine translation engine. As such, we think that adding more of these monolingual texts or possibly looking into a bilingual dictionary of financial terms may provide good results in the future.

TIMELINE AND COSTS

Based on our pilot, we estimate that training an MT engine for this purpose will take an additional three months before it can be deployed with reasonable translation quality. In the pilot, we used financial statements from the past 5 years for training. However, to fully train the engine, we recommend using all available data, which amounts to 10 years (or more, for some companies) of financial statements. Because auto-aligners yield poor results with the Japanese-English language pair, much of the time working on the MT engine will be spent in alignment. The rest of the time will be spent actually training the MT engine and then post-editing the translation, followed by a QA to check the progress and allow us to continually improve the engine.



An estimate of the costs associated with this project is as follows.

Task	Estimated Hours	Hourly Rate	Cost
Document alignment	40	$40.00	$1,600
MT training	12	$30.00	$360
Post-editing	8	$30.00	$240
QA	4	$30.00	$120
Total estimated project cost			$2,300

ANTICIPATED RESULTS

As noted above, the low BLEU score achieved in the pilot project was due mostly to the inability to access the source documents (only having access to the PDFs) as well as some of the PDFs we did have access to being password protected, limiting the amount of alignment we could do. However, we anticipate a machine translation that will achieve at least readability, which could then be post-edited, resulting in a high quality translation. Due to the time and labor saved from translating with a machine as opposed to human translation, we anticipate saving roughly 67% of the total costs of a full human translation and edit. While we do not anticipate a resulting translation that is to the same standard as that of a human translation, the goal is to get the information in a readable format to as many potential investors as possible, and this process will accomplish that goal, while saving significantly on costs.